2019

The effect of task complexity on rater severity in an adaptive performance-based second language oral communication test

Yongkook Won
Iowa State University

The effect of task complexity on rater severity
in an adaptive performance-based second language oral communication test

by

Yongkook Won

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Gary J. Ockey, Major Professor
Carol A. Chapelle
Elena Cotos
Amy G. Froelich
Volker H. Hegelheimer

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

DEDICATION

To my parents

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Gary Ockey, for his mentoring and advising throughout my PhD journey at Iowa State University. His classes and guidance opened my academic eyes to a new way of perceiving rater behavior in oral communication assessment. I also profoundly thank my committee members (in alphabetical order), Dr. Carol A. Chapelle, Dr. Elena Cotos, Dr. Amy G. Froelich, and Dr. Volker H. Hegelheimer for their guidance, support, and insightful feedback throughout the course of this project. I would also like to thank Dr. Gulbahar Beckett for giving me a chance to study IEOP assessment data.

I extend my sincerest thanks to my friends, colleagues, and the department faculty and staff, including the IEOP instructors, for making my time at Iowa State University a meaningful experience. My heartfelt thanks go to my professors, Dr. Young-kuk Jeong and Dr. Hyun-Sook Chung at the International Graduate School of English in Seoul and my colleagues at Yoons English School for their encouragement and support. I want to also offer my appreciation to my old friends (in alphabetical order), Haeyun, Hye-won, Hyunwoo, Jooyoung, Liberato, Mo, MoonYoung, and Phuong, including former and current members of the Language Assessment Student Organization, with whom I have enjoyed vivid academic discussions.

Finally, I would like to thank my brother and his family for staying with my parents while I have been away from home for my academic journey.

ABSTRACT

Despite the benefits of performance-based oral communication tests, a plethora of variables, as illustrated in Ockey and Li's (2015) model of oral communication assessment, can create construct-irrelevant variance in test scores. In relation to human participants in the oral communication tests, previous studies mostly focused on the direct effect of the rater group variable on test scores. Little attention has been paid to the interaction of raters with interviewers in oral communication tests. The present study investigates how raters evaluate test takers' performance in performance-based oral communication tests when interviewers can adaptively choose their questions, in terms of task complexity, responding to test takers' performance.

An explanatory sequential design with a mixed-methods approach was used to investigate the effect of task complexity on rater severity. For the initial quantitative data analysis, operational rating data from 1,689 test takers whose native languages are not English and scored by 24 certified raters in the Oral English Certification Test (OECT) and 162 audio recordings of 81 international graduate students in the OECT were analyzed with multilevel ordinal logistic regression, a paired samples *t*-test, and many-facet Rasch measurement (MFRM). To further investigate the effect of task complexity on rater severity, nine newly trained raters were trained to judge 80 speech samples of 40 test takers in the OECT. A partial credit model of MFRM was used to analyze raters' use of the scoring rubric depending on task complexity.

In the initial quantitative data analysis, *low* complexity prompts were statistically estimated as the most difficult item. The results of paired samples *t*-tests showed that only a few fluency measures demonstrated statistical differences by task complexity. The analysis

of the interaction of task complexity with rating contexts with nine newly trained raters using Welch's *t*-test showed that the difficulty of *high* complexity tasks decreased when raters became aware of the task complexity. This change of task difficulty suggests that raters in this adaptive performance-based oral communication may have changed their rating severity depending on their understanding of the task complexity. Follow-up verbal reports and interviews supported the findings in the quantitative analysis.

CHAPTER 1. INTRODUCTION

1.1. Context of the Problem

When it comes to measuring more productive language skills, especially speaking skills, performance-based assessments are typically more valuable in that they are able to elicit more construct-relevant language abilities than less direct test tasks, such as multiple-choice items, thus providing enhanced opportunities for generalizing test scores beyond the test context to the target language use (TLU) domain (Bachman & Palmer, 1996). Despite the benefits of performance-based assessments, there are also some challenges to overcome. A plethora of variables, in addition to a test taker's performance, can influence the scores in performance-based assessment (Ross, 2012), and the chances of having construct-irrelevant variance in the test scores increase accordingly. Ockey and Li's (2015) model of oral communication assessment illustrates that test takers, task types, technology, interviewers, rating scales, and raters in the given context can influence test scores. Factors in the oral communication model and their interactions with one another can create construct-irrelevant variance in the test scores, the components of which may create a situation where a certain group of test takers could receive an unfair advantage or disadvantage (AERA, APA, & NCME, 2014).

Among the variables, and aside from the test takers themselves, in the assessment model, score variability associated with the human participants (e.g., raters or interviewers in the oral performance interview) is one of the main causes of the construct-irrelevant variance (McNamara, 1996; Van Moere, 2006). As the performance in the oral communication model (e.g., oral proficiency interview; Liskin-Gasparro, 2003) is generally initiated by human interviewers and the quality of test takers' performance is measured by human raters who

have subjective judgments, construct-irrelevant variance by interviewers and/or raters is inevitable (McNamara, 1996). In addition, when raters must conduct an interview and rate at the same time, the chances of generating construct-irrelevant variance in the test scores would increase when raters are not adequately trained to perform both rater and interviewer roles.

Many studies have tried to identify the effects of raters and/or interviewers on the score variance of test takers in performance-based speaking ability tests (A. Brown, 1995, 2012; A. Brown & Hill, 1998; Chalhoub Deville & Wigglesworth, 2005; Karavas & Delieza, 2009; Van Moere, 2013). These studies investigated how construct-irrelevant variables, such as raters' first language (Kobayashi, 1992; Zhang & Elder, 2010), raters' experience (Davis, 2016), rater training (Davis, 2016), raters' familiarity with certain pronunciation varieties (Carey et al., 2011; Yan, 2014), and gender (A. Brown & McNamara, 2004; O'Sullivan, 2000), affected test scores. Most of these studies, however, seem to assume that performance-based oral communication tests are static: that is, test takers and interviewers are guided to follow pre-determined conversation paths. This assumption could be true if the test were administered using pre-recorded prompts, as in computer-based, semi-direct oral performance tests (Kim, 2015). Even though the number of tests delivered by computers is increasing (Chapelle & Douglas, 2006), interview-based oral communication proficiency tests could not easily be replaced by computer-based tests until new technologies are able to elicit the variety of performances, as human interviewers do, in the context of interest. Thus, the performance-based oral communication test should be dynamic in terms of the interactions among test participants (i.e., test takers and interviewers) when the construct of the test concerns the off-line nature of the TLU tasks.

Regarding the dynamics of performance-based oral communication tests, human interviewers likely adjust their interactions with test takers and these adjustments should also be considered when defining the construct of interest for users of such tests. If test takers do not understand the initial prompt, they are unable to produce appropriate responses. Accordingly, it may not be possible for interviewers to elicit ratable samples from some test takers without simplifying the language or slowing down the speech rate of the question. Previous literature has focused on the features of interviewer behaviors that lead to rating compensation (A. Brown, 2005; Ross, 1992); however, no studies have focused on the effects of interviewers' adaptive behaviors depending on the task difficulty and the effect of adaptive behaviors on the evaluation of test takers' performance. Nakatsuhara (2011) investigated the interactiveness among test takers and interviewers, but focused more on the relationship between test takers' listening proficiency and their speaking performance; even though she studied the dynamics of communications on the part of test takers (e.g., test takers' asking for repeating or rephrasing of the questions when they fail to understand them), the interviewers were not allowed to dynamically interact with the test takers by changing their prompts under the IELTS test setting used in her study. These studies that assumed the non-adaptiveness of interviewers would not be sufficient for analyzing the effects of interviewers and/or raters in performance-based oral communication tests, nor would provide potential solutions to mitigate any bias associated with interviewers and/or raters. Hence, research on interviewers' adaptiveness in performance-based tests and its effect on the performance of the test takers and scoring by raters needs to be conducted to better understand and interpret the scores of performance-based oral communication tests.

1.2. Purpose of the Study

This study investigates how interviewers of a performance-based oral communication test behave in an adaptive context, where task complexity should be adjusted to the test takers' oral communication proficiency, and how raters of test takers' performance interpret the evidence when interviewers adapt their roles depending on test takers' behaviors. These behaviors are based on a consideration of other variables in the oral communication assessment model (Ockey & Li, 2015). Unlike the current research trends in oral proficiency interview assessments, which are designed to increase the reliability of the test by using pre-determined automatic prompts that constrain the variability of interviewers (Ross, 2012), this study investigates the characteristics of the factors in the oral communication assessment model when interviewers' adaptive nature, or the interviewer's natural interventions, is preserved.

Those studies on the effects of interviewers or raters in performance-based oral communication tests have been conducted using both quantitative and qualitative methods. Studies employing quantitative analysis (McNamara & Lumley, 1997) generally have treated interviewers and/or raters as facets in which individual interviewer or rater behaviors are consistent, irrespective of task types or test takers. In terms of interviewers' natural language use in the oral communication test, however, interviewer behaviors can be sensitive and reactive to test takers' performance, and interactions among interviewers and test takers can vary throughout the test session. Thus, aggregating the effects of interviewers or raters on test taker performance through the entire test session would yield a loss of some valuable information that may explain the variance of test scores. Studies using qualitative analyses (A. Brown, 2003; A. Brown & Hill, 1998; Lazaraton, 1996), on the other hand, generally

have investigated language use with conversation analysis. These types of studies can help us understand the linguistic features used in the tasks, but cannot fully support the generalization of observed scores beyond the test context to other situations.

This study focuses on the interactions between test takers and interviewers in terms of task complexity and its effect on test takers' performance scores rated by human raters in an oral proficiency interview for international teaching assistants at a United States (U.S.) university.

1.3. Research Questions

To investigate the variation in raters' score assignment in relation to task complexity in an adaptive performance-based oral communication test and its effects on test scores, two main research questions (RQs) are addressed. The first main research question (RQ 1) is "how do raters adjust their score assignment depending on the complexity level of the prompts in an adaptive performance-based oral communication test?" and the second research question (RQ 2) is "how do raters adjust their score assignment depending on their understanding of prompt complexity in a performance-based oral communication test?" As the first main research question (RQ 1) presupposes that interviewers select prompts with different complexity depending on test takers' performance, interviewers' adaptive selection of task prompts is firstly addressed with the first sub-research question (RQ 1-1). The effect of task complexity on test takers' performance (RQ 1-2 and RQ 1-3) is then adopted to sort out the effect of task complexity on test takers' performance from raters' score assignment variation depending on task complexity. Raters' adjustment of their score assignment is finally regarded with the last sub-question (RQ 1-4) about change in rating severity affected

by task complexity. The second research question (RQ 2) is more deeply explored by investigating raters' scoring behaviors in an experimental condition. In addition to the first sub-question (RQ 2-1), which attends to raters' adjustment of score assignment, two additional sub-questions (RQ 2-2 and RQ 2-3) are posed to fully understand how raters adjust their score assignment depending on their understanding of prompt complexity. The following research questions are addressed in the current study:

RQ 1. How do raters adjust their score assignment depending on the complexity level of the prompts in an adaptive performance-based oral communication test?

RQ 1-1. To what extent does the performance (in holistic scores) of test takers in each task affect interviewers' selection of the complexity level of the following tasks in an adaptive performance-based oral communication test?

RQ 1-2. How does task complexity determined by the interviewer in performance-based L2 oral communication tests affect test takers' oral output in terms of their linguistic complexity and fluency measures?

RQ 1-3. How does task complexity determined by the interviewer in performance-based L2 oral communication tests affect test takers' proficiency scores (lexico-grammar, and fluency) when graded by human raters?

RQ 1-4. How do raters change their rating severity depending on the complexity level of the prompts in an adaptive performance-based oral communication test?

RQ 2. How do raters adjust their score assignment depending on their understanding of prompt complexity in a performance-based oral communication test?

RQ 2-1. To what extent do raters change their rating severity depending on their understanding of prompt complexity in a performance-based oral communication test?

RQ 2-2. To which evaluation categories do raters attend together with task complexity while scoring oral communication audio clips?

RQ 2-3. How do raters apply task complexity to their interpretation of evaluation criteria in terms of rating severity?

To answer RQ 1-1, an ordinal logistic regression model with a holistic score of the preceding task as a predictor variable and selection of the following task complexity as a dependent variable is used. To answer RQs 1-2 and 1-3, multiple paired samples *t*-test (or non-parametric equivalent measures) are used to compare the effect of task complexity on linguistic outputs and the performance scores. To answer RQ 1-4, a partial credit model (PCM) (Wright & Masters, 1982) of the data with many-facet Rasch measurement (MFRM) is used to analyze raters' score assignment depending on task complexity. To answer RQ 2-1, another PCM model is used to analyze raters' behaviors in a more controlled experimental condition. To answer RQs 2-2 and 2-3, in addition to the statistical score analysis, raters' retrospective verbal reports and interviews are used to identify how raters assign the scores in performance-based oral communication tests. Detailed data analytic methods for each research question are described in the Methodology section.

1.4. Significance of the Study

From a theoretical point of view, this study of interviewer dynamics in the scoring of performance-based oral communication tests in an adaptive context provides information on what factors influence raters' cognitive processes when they assign scores against given evaluation rubrics. More specifically, this study provides insights into the way raters respond to interviewers' behavior, in terms of task complexity, in a performance-based oral communication test; thus, the conceptual mechanisms of interviewers' choice of task prompts with various complexity and raters' interpretation and application of the scoring rubrics into rating processes can be better understood.

From a practical standpoint, the understanding of interviewers' adaptive selection of the task prompts in a performance-based oral communication test enhances our understanding of potential construct-irrelevant variance and helps to develop rater and interviewer training guidelines that can help raters assign fairer test scores. A detailed recognition of the variabilities of scoring with the given scale, coupled with interviewers' selection of task prompts, can supply more in-depth feedback to test takers. As the relationship between interviewers' selection of task complexity and raters' use of rating scales is evaluated, the findings in this study help raters understand to which rating characteristics they should attend when task complexity fluctuates in an L2 oral communication test.

CHAPTER 2. LITERATURE REVIEW

This chapter first describes the Oral English Certification Test (OECT) for prospective international teaching assistants (ITAs), who are non-native speakers of English. The general description of the ITA test is followed by a description of the oral English proficiency test from which the data for this study are derived. In addition, the model of oral communication assessment (Ockey & Li, 2015) used in this study is introduced with a review of the literature on score variance in the oral communication tests in terms of both construct-relevant and irrelevant variance. The effect of the task prompt in performance-based oral communication assessment is also discussed to support the explanation of interviewer and rater behaviors during the assessment process. A review of linguistic measures is also introduced to help understand how task complexity affected test takers' linguistic performance. This chapter ends with descriptions of the statistical procedures used in associated data analysis.

2.1. Oral English Assessment for ITAs

As the number of international teaching assistants (ITAs) who fill the teaching assistant role in undergraduate courses in U.S. universities have increased, having qualified ITAs has become an integral part of education at the post-secondary level (Ginther, 2003). The concern about ITAs' English communication skills, therefore, has increased correspondingly (Bailey, 1983; Farnsworth, 2013). For this reason, many universities with a large number of ITAs have implemented ITA-specific language assessments to screen out unqualified teaching assistants or to identify ITAs in need of further oral English training (Ginther, 2003). English oral communication proficiency tests for prospective ITAs generally

consist of a conversational interview, in which ITAs are asked to discuss general academic topics, and a formal classroom teaching simulation, in which ITAs present topics in their own disciplines (Douglas, 2000; Ginther, 2003). The ITA-specific language tests are based on ITAs' performance and are designed to elicit ratable samples of spoken English from the ITAs.

The English oral communication proficiency test for ITAs is high-stakes in that resulting test scores are used for the departments of the test takers to make informed decisions when assigning teaching duties to the test takers. In addition, test scores are used to assign ITAs to appropriate English oral communication classes when they fail to meet the required levels of English proficiency in their departments. If there are false positives, or ITAs get higher scores than what actually represents their true abilities, undergraduate students taking courses with ITAs would encounter a poor educational environment with unqualified teaching assistants in their classroom. On the other hand, if there are false negatives, or ITAs receive lower scores than they should based on their true abilities, the ITAs could take additional unnecessary courses and universities would be required to spend unnecessary resources for training ITAs. Thus, it is imperative that the reported scores for English oral communication proficiency tests for ITAs be reflective of their targeted English oral communication abilities.

2.2. Oral English Certification Test

The Oral English Certification Test (OECT), which was developed in 1985 and is administered by Iowa State University (ISU) (Douglas, 2000), is a performance-based oral communication test that assesses how effectively ITAs can communicate in English in

university life or classroom teaching situations. The OECT consists of two sections: the Oral Proficiency Interview (OPI) and the Teach Simulation (TEACH) (Cotos, 2014). The OPI aims to measure how well ITAs use oral English in conversation-based contexts, including conversations with their professors, their students, and university staff. The TEACH aims to measure how well ITAs perform on the instruction of teaching topics in their classroom teaching contexts. The evaluation rubrics of both the OPI and TEACH sections include four components: *functional competency*, *fluency*, *lexico-grammar*, and *pronunciation* (see Appendix A for the rating scale).

The OPI consists of four tasks, three impromptu oral communication tasks and one role-play task, and takes test takers approximately 10 minutes to complete. Before the four main tasks, the test starts with a one-minute warm-up conversation between an interviewer, who is one of the three raters, and a test taker. The oral production of a test taker in the warm-up part is not graded, but based on the test taker's performance in the warm-up task, an interviewer selects an appropriate level task out of five levels of tasks (1, 2, 3+, 3-, and 4) for the following scored task. The impromptu oral communication tasks require test takers to respond for two minutes without any preparation time, while the role-play task gives one minute of preparation before the two-minute role-play with the interviewer. Three raters, including a rater, who also served as an interviewer, assign scores on an 18-point scale (13-30) with four proficiency levels by using an online evaluation platform. The scores in the four tasks are averaged to create a final OPI section score for the test taker. In addition, each rater gives a separate general impression score on the test taker's OPI performance. In the online evaluation platform of the OECT, therefore, each test taker receives three sets of four task scores and a general impression score by each of the three raters.

After completing the OPI section of the OECT, test takers enter into the TEACH section. In the TEACH portion, test takers are given one hour of preparation time in a quiet room with textbook materials on the TEACH topic on which they need to conduct a teaching demonstration. Test takers choose one out of approximately 10 undergraduate introductory topics in their field of study for the TEACH demonstration. During the TEACH test, test takers are given two minutes to prepare and write teaching notes on the board. Test takers simulate teaching the topic they have chosen for five minutes in front of three raters, followed by a three-minute question and answer session. Test takers are asked to simulate what they would do in a real-life classroom context and raters act as students and ask questions as if they are in class. As the TEACH section lasts only 10 minutes, including two minutes of writing on the board, test takers generally explain only one concept of their chosen topic. As in the OPI section, three raters assign scores on the 18-point scale along four proficiency levels. However, there are no sub-tasks in the TEACH section; thus, each rater assigns a holistic score for the overall language effectiveness and gives comments on the test taker's performance and optionally gives comments on the diagnostic features, such as comprehensibility, pronunciation, fluency, vocabulary, grammar, pragmatics, and listening skills. There are additional rating criteria regarding cultural ability, which assess how well test takers handle the classroom situation, such as maintaining eye contact, developing a rapport with the class, demonstrating familiarity with the cultural code, and appropriately using chalkboard.

2.3. Performance-Based Oral Communication Assessment Model

A model of oral communication assessment proposed by Ockey and Li (2015) presents many of the factors that are involved in the interpretation of the test scores of performance-based oral communication tests, as shown in Figure *2.1*. According to Ockey and Li's model, test scores in performance-based oral communication tests are assigned based on raters' evaluation of the test takers' oral performance on the given tasks with respect to the rating scales, interviewers' personal characteristics, and technology in the given context. The arrows in the model indicate the direction of the influence among factors, and the factors have either direct or indirect influence on test scores.

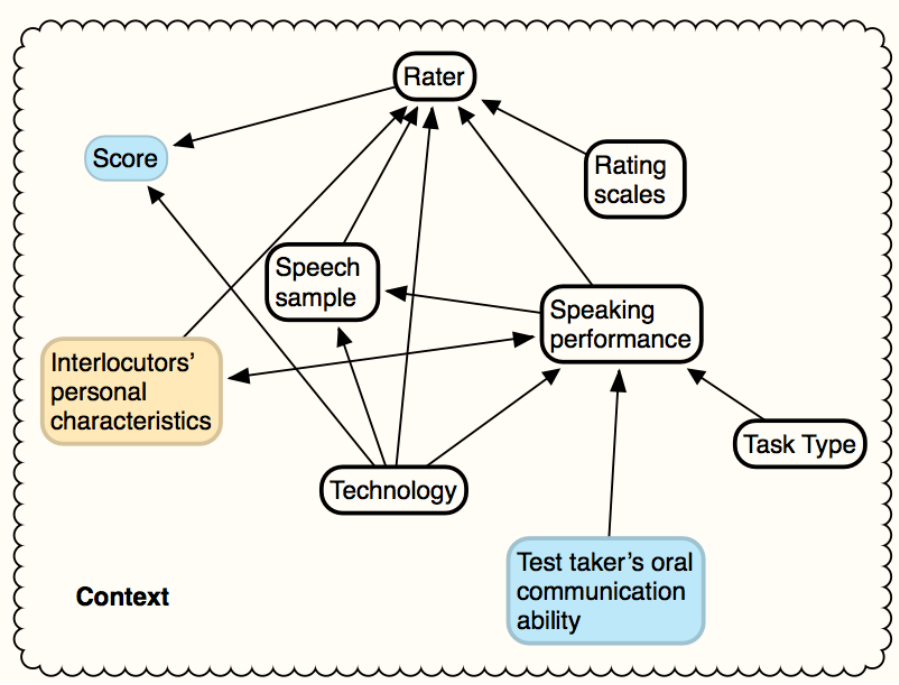


Figure 2.1. A model of assessment of oral communication (adapted from Ockey & Li, 2015, p. 2)

Test scores represent attributes of test takers, which should be the stated construct of a test, and are reflected in test performances (Cronbach & Meehl, 1955). According to the

interactionalist perspective, one that is theoretically well supported in language testing (Chapelle, 1998), test scores in performance-based speaking tests are indicators of test takers' oral communication abilities in the given context. As many factors are involved in the interpretation of test scores, there is likely to be undesired construct-irrelevant variance in performance-based oral communication test scores. The construct-irrelevant variance can be generated from any element in the model, such as rating scales, task types, interviewers' personal characteristics, technology used in the given context, or their interactions (Ockey & Li, 2015). Among the factors in the oral communication assessment model, human variables (i.e., raters and interviewers) seem to be one of the most influential, but not easily detected, factors in contributing to the construct-irrelevant variance; this is because other factors are generally fixed before the administration of the test, while raters and interviewers act adaptively depending on the performance of the test takers and the task types.

2.4. Raters and Interviewers in Performance-Based Oral Assessments

Even though the variability in test scores could be generated due to (a) test takers' communication ability, (b) raters' behaviors, and (c) tasks and other environments (McNamara, 1996), the scores are assigned based on raters' interpretation of test samples against the evaluation criteria (Van Moere, 2013). This means that the sources of score variation should be perceived and interpreted by raters during the assessment process. Thus, it is important to minimize the score variabilities associated with rater behaviors, or construct-irrelevant facets in the test, to increase the validity of test scores. According to McNamara (1996), the score variability associated with raters originates from (a) raters' overall severity, (b) raters' systematic patterns of leniency toward a group of candidates or

particular tasks, (c) raters' differential interpretation of rating scales, and (d) raters' random inconsistency in scoring.

Most of the studies on rater effects have focused on raters whose role is rating on-site or rating using video/audio recordings (A. Brown, 1995; Chalhoub Deville & Wigglesworth, 2005; Kim, 2011), but not many have focused on the raters who also serve as an interviewer in the performance-based oral speaking test. Wolfe (2004) summarized that most studies on rater effects have focused on raters' cognitive processing, characteristics of raters, rating tasks, rating environment, and the impact of rater effects on rating. These studies mostly investigated raters who assess the performance of test takers independent of playing a role as an interviewer to elicit the performance from test takers in the oral proficiency interview.

Another line of studies on rater effects includes the interaction among interviewers, test takers, and raters. Test takers' performance should be measured consistently, no matter who scores their performance or who interviews them (A. Brown, 2012), but construct-irrelevant score variance related to raters and interviewers has been reported in the studies of the performance-based assessments (A. Brown, 2003; McNamara & Lumley, 1997). Some studies have examined the influence of the interviewers' backgrounds, such as gender (O'Loughlin, 2002, 2007; O'Sullivan, 2000), language proficiency (Davis, 2009), and ethnicity (Hou, 2006). Other studies investigated the relationship between score variability and interviewers' competency (Morton, Wigglesworth, & Wiliams, 1997), areas of concern (May, 2011), and helpfulness (Chartrand & Bargh, 1999; Clark, 2002; Wilson & Wilson, 2005). Some studies found that when interviewers were more supportive, test takers were more likely to show their true ability and earn higher scores (McNamara & Lumley, 1997; Morton et al., 1997), but other studies reported that test takers with less supportive

interviewers received better scores, because raters compensated for the lack of support by the interviewers (Lazaraton, 1996).

2.5. Effects of Prompts/Tasks in Performance-Based Oral Communication Assessments

Scores in performance-based oral communication tests are mainly affected by test takers' oral communication proficiency, rater severity, and task difficulty (Fulcher, 2003), in addition to other factors in the oral communication assessment model (Ockey & Li, 2015). As different tasks elicit different speech samples in terms of their complexity, accuracy, and fluency (Ellis, 2009; Robinson, 2011b; Skehan, 2009), the varying task types used in oral communication assessments could influence the scores of the same test (G. Brown, Anderson, Shillcock, & Yule, 1984). Hence, to better understand test scores and fairly interpret them, it is important to understand how task characteristics (Chalhoub-Deville, 1995; Skehan & Foster, 1997; Upshur & Turner, 1999), including task difficulty, affect test scores and the language elicited.

<table>
<tr><td colspan="4">Degree of difficulty
(low) ----------→ (high)</td></tr>
<tr><td colspan="2">Static task</td><td>Dynamic task</td><td>Abstract task</td></tr>
<tr><td colspan="2">e.g., diagram, pegboard</td><td>e.g., story, car crash</td><td>e.g., opinion</td></tr>
<tr><td rowspan="2">Degree of difficulty</td><td colspan="3">Many elements, relationships, characters, etc.
(more difficulty)</td></tr>
<tr><td colspan="3">Few elements, relationships, etc.
(less difficulty)</td></tr>
</table>

Figure 2.2. Tasks of ascending difficulty (G. Brown et al., 1984, p. 64)

Many studies have proposed criteria to be considered when categorizing the degree of task difficulty or complexity in second language performance. For example, G. Brown et al. (1984) proposed a framework of task difficulty, as shown in Figure *2.2*. According to G. Brown et al.'s model, task difficulty increases from static tasks (e.g., diagram description) to

dynamic tasks (e.g., storytelling, car crash description), and to abstract tasks (e.g., opinion giving) coupled with the amounts of information, relationships, and characters that needs to be covered.

Prabhu (1987) provided another criterion for grading tasks with the measures of cognitive complexity. In his model, Prabhu suggested the following five parameters to be considered in analyzing task difficulty: (a) the amount of information to be handled in the task (*information provided*), (b) the distance between the given information and the target information (*reasoning needed*), (c) how precise the information should be interpreted for the given tasks (*precision needed*), (d) the degree of learners' familiarity with the tasks (*familiarity with constraints*), and (e) the degree of abstractness of the tasks.

Among the taxonomies proposed to categorize the task characteristics in applied linguistics, most frequently cited models are Skehan's (1998, 2001) *limited capacity model* and Robinson's (2001, 2011b) *triadic componential framework* (Révész & Gurzynski-Weiss, 2016). Skehan proposed three categories to assess L2 task difficulty: *code complexity*, *cognitive complexity*, and *communicative stress*. *Code complexity* refers to the linguistic demands on learners or test takers in the tasks. *Cognitive complexity* consists of cognitive familiarity and cognitive processing. Cognitive familiarity is concerned with topic, discourse genre, and task familiarity, while cognitive processing denotes how much cognitive effort test takers should exert when completing the tasks. *Communicative stress* is related to the stress that test takers would have during the test, such as time limits or the type of response. In his meta-analysis of the task-based performance studies, Skehan (2001) suggested that familiarity of information, discourse style (dialogic versus monologic), degree of sequencing structure, complex outcomes, and transformation requirement would influence the task

difficulty. Skehan's limited capacity model has been supported by the literature (Weir, O'Sullivan, & Horal, 2006).

Robinson's (2001, 2011b) *triadic componential framework*, which analyzes the complexity of tasks that influence the task-based performance, distinguishes *task complexity*, *task difficulty*, and *task conditions*. As illustrated in Figure *2.3*, *task complexity* is concerned with the cognitive demands of the task, which is similar to Skehan's concept of cognitive complexity (Révész & Gurzynski-Weiss, 2016). An increased number of elements in the tasks or multiple concurrent tasks would make the task more difficult than a simple single task. According to Robinson's framework, *task difficulty*, which Skehan (1998) used as an umbrella term to include code complexity, cognitive complexity, and communication stress (Révész & Gurzynski-Weiss, 2016), is related to learner factors. If learners are motivated or have confidence, the difficulty would be perceived as easier than when they are less motivated and lack confidence. *Task conditions* are the interactive demands of the task, such as communication direction, gender of the communication partners (e.g., interviewers, peers in group oral tests), or the power relationship among the partners (e.g., teacher-student relationship).

Task complexity (cognitive factors)	**Task conditions (interactive factors)**	**Task difficulty (learner factors)**
a) resource-directing e.g., ± few elements, ± here-and-now, ± reasoning demand	a) participation variables e.g., one-way/two-way, convergent/divergent, open/closed	a) affective variables e.g., motivation, anxiety, confidence
b) resource-depleting e.g., ± planning time, ± single task, ± prior knowledge	b) participant variables e.g., gender, familiarity, power/solidarity	b) ability variables e.g., aptitude, proficiency, intelligence
Sequencing criteria Prospective decisions about task units	Methodological criteria On-line decisions about pairs and groups	

Figure 2.3. A triad of task complexity, conditions, and difficulty factors (Robinson, 2001, p. 294).

In the current study, *task complexity* and *task difficulty* in Robinson's *tragic componential framework* are used to indicate different aspects of the task characteristics, because this study investigates how the task characteristics influence test takers' performances and how raters perceive the performance as assessed via the evaluation criteria. *Task complexity*, following Robinson's definition, is used to indicate the *cognitive complexity* of the task itself. On the other hand, *task difficulty*, in Robinson's model, is used to indicate how test takers perceive the difficulty of the prompts. *Task conditions* in Robinson's model is not considered in the current study, because the task conditions in the oral proficiency interview can be considered nearly equivalent across all task prompts. Skehan's umbrella term *task difficulty* is not used, because *cognitive complexity* and *communicative stress* in Skehan's model do not vary across the prompts used in the current study. As *cognitive complexity* in Skehan's model is mainly related to task familiarity and the interview prompts that are used in the current study are the questions that test takers generally face in everyday life, it can be assumed that the *cognitive complexity* is almost equivalent across the prompts. In addition, *communicative stress* in Skehan's model is not considered in the current study, because this study statistically investigates the *task difficulty* in Robinson's term that partially covers the concept of *communicative stress*, a feature not easily measured during a test.

The models categorizing the tasks are originally developed for the syllabus design of task-based language learning, and task variation does not necessarily bring score variations in performance-based oral communication tests. For example, Leaper and Riazi (2014) found that when the scores in a test are represented by a single holistic score, the test taker's score cannot represent the varieties of linguistic differences. With the same scores, the content of the speech sample could be varied. In this vein, task variation would not necessarily create

score differences in some instances (Fulcher & Reiter, 2003). The point would be that test scores represent how raters perceive and assess the utterances by test takers coupled with other variables in the test. Thus, it is important to understand how task complexity interacts with test takers, interviewers, and raters in the performance-based assessments and to minimize any interactions that would cause construct-irrelevant variance in the test scores.

2.6. Review of Linguistic Measures

To understand the relationship between proficiency scores graded by human raters and linguistic features in the spoken data depending on the task complexity level, the speech data need to be linguistically analyzed with analytic scoring criteria in the evaluation rubric: *functional competency*, *fluency*, *lexico-grammar*, and *fluency*. However, *functional competency* and *pronunciation* scoring are not discussed in the linguistic analysis section of the current study. This is because *functional competency* can be subjectively measured by raters and it is difficult to quantify with linguistic indices, and *pronunciation* is rarely believed to be affected by task complexity.

To understand the *lexico-grammar* scoring, two measures of linguistic complexity (*Subordinate Index* and *Guiraud Advanced 1000*) and one measure of linguistic accuracy (*Errors per AS-unit*) were chosen. First, the *Subordination Index* (Beaman, 1984) measures the syntactic complexity of the speech, because the index shows how well test takers use complex syntax (Michel, 2011). The *Subordination Index* indicates the ratio of subordinate clauses that work as the sentence subject, verb complement, or phrasal post-modifier, per AS-unit. Second, the *Guiraud Advanced Index* (Daller, Van Hout, & Treffers-Daller, 2003) measures the lexical complexity instead of the commonly used *Type-Token Ratio (TTR),*

because TTR fails to consider the length of the text (Hout & Vermeer, 2007). In addition, the *Guiraud Advanced Index* holds advantages over the simple *Guiraud Index* in that every word carries a different weight when it is perceived by human raters (Daller et al., 2003; Hout & Vermeer, 2007). The *Guiraud Advanced 1000 Index* is calculated by dividing the number of advanced word types (1,000 frequency level in the current study) by the square root of the number of tokens. The 1,000 most frequent word families in the British National Corpus (BNC) (Leech, Rayson, & Wilson, 2001) are used as criteria to decide the advanced words in the *Guiraud Advanced 1000 Index* in the current study. The 2,000 most frequent word families in English are not used, because they cover more than 85% of the running words, or tokens, in the BNC (Nation, 2006); also, the *Guiraud Advanced Index* with the 2,000-frequency level does not produce any meaningful numbers considering the length of the speech files in the current study are short and are produced by non-native speakers who are likely to have a smaller vocabulary size than native speakers of English.

With respect to *fluency*, temporal measures (*Speech Rate*, *Mean Length of Utterance*, and *Phonation-Time Ratio*) and dysfluency markers (*Repairs per AS-unit*, *Filled Pauses per AS-unit*, and *Preparation Time*) are used (Lennon, 2006). The *Unpruned Speech Rate* is calculated by dividing the number of syllables by the amount of total time used, as recommended by Riggenbach (1991). The *Pruned Speech Rate* is also calculated by using the syllables after having cleaned for repetitions, repairs, and reformulations. The *Mean Length of Utterance* is calculated by averaging the number of syllables between pauses of 0.25 seconds or more, because pauses above 0.25 seconds are considered as the most reliable cut-off points of runs (Towell, Hawkins, & Bazergui, 1996). The *Phonation-Time Ratio* is the "percentage of time spent speaking as a percentage proportion of the time taken to produce

the speech sample" (Towell, 2002, p. 120) and is automatically calculated using the Praat script (De Jong & Wempe, 2009). As for dysfluency markers, *Repairs per AS-unit* are calculated by counting the number of repairs (e.g., repetitions, reformulations) per AS-unit; *Filled Pauses per AS-unit* are calculated by counting the number of filled pauses (e.g., uh, uhm) per AS-unit; and *Preparation Time* is measured by the length of time in seconds between the end of interviewers' utterances and the beginning of test takers' utterances.

2.7. Review of Analysis Methods

Explanatory designs in a mixed-methods approach (Mackey & Gass, 2016) are used in the current study. The quantitative data from the OPI of the OECT are analyzed before collecting the qualitative data (e.g., rater interview and linguistic analysis). The interview and verbal reports data are used to better understand the findings in the quantitative data analysis. Four analytic methods are used: multilevel ordinal logistic regression analysis, paired samples *t*-test (or Wilcoxson signed-rank test), many-facet Rasch measurement analysis, and retrospective verbal report and interview analysis.

2.7.1. Inter-coder Reliability

Inter-coder reliability, or inter-rater reliability, is defined as "the extent to which two or more independent coders agree on the coding of the content of interest with an application of the same coding scheme" (Lavrakas, 2008, p. 344). Among numerous different statistical indices of inter-coder reliability, only several indices are widely used in applied linguistics studies: *percent agreement, Cohen's kappa, Cronbach's Alpha, and Krippendorff's alpha*. *Percent agreement* is the proportion of the agreed upon codes by two observers. *Percent*

agreement is limited to nominal coding with only two coders with the same number of coding categories. In addition, the *percent agreement* does not account for the coding agreement by chance. *Cohen's kappa* (J. Cohen, 1960) is another widely used inter-coder reliability index for the nominal coding, which takes into account the agreement that can occur by chance. Even though *percent agreement* and *Cohen's kappa* are widely found in applied linguistics journals, both *percent agreement* and *Cohen's kappa* can only be used with two coders. In contrast, *Cronbach's alpha* (Cronbach, 1951) and *Krippendorff's alpha* (Krippendorff, 1970, 2004a) can be used with multiple coders. *Cronbach's alpha* can measure the consistency of two or more observers, and it is also used for inter-coder reliability in applied linguistics. *Cronbach's alpha* is a statistic for interval and ratio-level data and compares the sum of coders' variance with the variance of total test scores; that is, *Cronbach's alpha* measures the covariation of the item scores, but does not measure the agreement among coders. On the other hand, *Krippendorff's alpha* is a statistic for nominal, ordinal, interval, and ratio-level data, and it compares the observed disagreement and expected disagreement of the coders. For these reasons, *Krippendorff's alpha* may be the most suitable inter-coder reliability index in a study in which the agreement of coding with different levels of data (i.e., nominal, ordinal, interval and ratio) by combinations of two or more coders is investigated. The suggested benchmarks for *Krippendorff's alpha* values are as follows: values equal to or greater than 0.80 is adequate and values between 0.67 and 0.80 are acceptable (Krippendorff, 2004b).

2.7.2. Multilevel Ordinal Logistic Regression

Logistic regression is a special case of regression model in which the response variables are categorical data (Snijders & Bosker, 2012). Like ordinary least squares (OLS) regression, logistic regression is an approach to prediction; however, in OLS regression, the expected values of the response variable are modeled with a function of predictor variables, while in logistic regression, the probability or odds of the response variable taking a particular value is modeled through the application of logit-link (Agresti, 2007; O'Connell, 2006). The logit is the natural log of the odds where the odds for an event indicates the probability of the success of the event to that of the failure. Ordinal logistic regression is one of the logistic regression models where the outcome variable is ordinal data and the explanatory variables are discrete and/or continuous data (Agresti, 2007; O'Connell, 2006). Ordinal logistic regression is different from multinomial logistic regression in that ordinal logistic regression can compare the probability of getting a certain response category when taking into account the ordering.

Multilevel ordinal regression is an ordinal regression where the response variable is nested within group variables. Prior to using multilevel ordinal regression, intraclass correlation coefficient (ICC) is calculated to check the variability of between-group variables. If ICC is large, multilevel analysis should be used. ICC for the ordinal regression is calculated as follows: $\frac{between-group\ variance}{between-group\ variance\ +\text{Level}-1\ \text{residual}}$. Level-1 residuals for the ordinal regression are assumed to follow the standard logit distribution, with a mean of zero and a variance of $\pi^2/3 = 3.29$ (Snijders & Bosker, 2012). Even when ICC is small, it is still recommended to use multilevel analysis if the data has a multilevel structure (Nezlek, 2008). Multilevel ordinal regression requires three assumptions. First, the dependent variable should

be ordinal data. Second, there should be no multicollinearity among independent variables. Finally, the proportional odds assumption should be met. The proportional odds assumption, or parallel lines assumption, is the effect of an independent variable on the ordinal dependent variable is uniform over all of the levels of the dependent variable (O'Connell, 2006).

2.7.3. Paired Samples *t*-test and Wilcoxon Signed-rank Test

The paired samples *t*-test, or the dependent *t*-test, is a statistical procedure used to compare the mean scores in the two conditions (Field, 2009). In a paired samples *t*-test, each subject is measured twice. The paired samples *t*-test assumes that (a) the sampling distribution of the difference between two paired scores should be normal, (b) the score difference data should have no significant outliers, and (c) the dependent variable should be continuous data (Field, 2009). The Wilcoxon signed-rank test (Wilcoxon, 1945) is a non-parametric alternative to the paired samples *t*-test. The difference between the paired samples *t*-test and the Wilcoxon signed-rank test is that the paired samples *t*-test is based on the score differences, while the Wilcoxon signed-rank test is based on the rank differences. When the assumption of the paired samples *t*-test is violated, the Wilcoxon signed-rank test should be used.

2.7.4. Many-Facet Rasch Measurement (MFRM) Analysis

Rasch models, one of the item response theory (IRT) models, are probabilistic measurement models that can calibrate parameters (e.g., test taker ability, item difficulty) independently of each other (Bond & Fox, 2015; McNamara, 1996). Rasch models use only item difficulty out of three parameters (i.e., difficulty, discrimination, and pseudo-guessing)

in the IRT models. Unlike classical test theory, the parameter estimation in Rasch (or IRT) is sample-independent (de Ayala, 2009). Rasch models are able to capture test takers' ability (and other relevant facets) when estimating the difficulty of the items. Thus, the item difficulty does not change depending on test taker ability. Ability estimation is, therefore, said to be item-independent. As Rasch calculates the parameters in the model beyond the level of the sample used, the calculated parameters are generalizable to the population if the model fits well.

The latent variables (e.g., test taker ability or item difficulty) in Rasch models are expressed in *logit* scale, which is the natural logarithm of the odds ratio and ranges from negative infinity to infinity (Bond & Fox, 2015; de Ayala, 2009). The variables in Rasch models are assumed to be located on the same item-person logit scale continuum. Any test taker on an item located at the same point is likely to have a 50% chance of getting the item correct if it is a dichotomous item or a 50% chance of getting the given item score if it is an item with a continuous rating scale (Bond & Fox, 2015; McNamara, 1996). If the item difficulty is one logit higher (or lower) than the person's ability level, the chances of success on the item increase (or decrease) to about 75% (or 25%). If the difference is two logit levels, the chances of success increase (or decrease) to about 90% (or 10%) (Bond & Fox, 2015). As the logit scale is linear in the parameters, the effects of predictor variables on logits are additive (O'Connell, 2006). In addition, as the variables are expressed on the true interval scale, raters' use of rating scales in performance-based language assessments can be easily compared.

Many-facet Rasch measurement (MFRM) is an extended version of the Rasch model and can simultaneously calibrate more than two variables (or facets) that can influence

assessment outcomes (Eckes, 2015). The original Rasch model, proposed by Rasch (1960), illustrates that the probability of getting a correct answer is a function of the difference between only two facets (i.e., test taker ability and item difficulty), while MFRM considers more than two facets at the same time (i.e., test taker ability, item difficulty, rater severity, test form) (Bond & Fox, 2015). In comparison with G-theory (Shavelson & Webb, 1991), which can also analyze data with multiple facets, MFRM provides information about the relationship among facets, such as test taker abilities, task difficulty, and rater severity, and estimates their relative locations in a linear scale. These individual scores would be useful to help explain the adaptiveness of the participants in the current study. MFRM can identify individual sources of the interactions among facets and provide a map of the distribution of each facet.

MFRM follows the following assumptions: local independence of items, unidimensionality, absence of guessing, and equal item discrimination (Bond & Fox, 2015). First, the local independence of items can be measured by using Rasch-Cohen's Kappa and investigating the rating context. Second, the unidimensionality assumption of the construct in the test can be checked by investigating infit and outfit mean-square values. Third, there should be no *guessing* in the data. Finally, all items are assumed to have the same *discrimination*.

2.7.5. Retrospective Verbal Report Analysis

Verbal reports are study participants' comments about their cognitive processes during the tasks they are required to complete (Bowles, 2010). In language testing, verbal report methodology has been used to support the validity of language tests or to examine test-

taking strategies (Bowles, 2010). A. Cohen (1998) outlined three main types of verbal reports: self-report, self-observation, and self-revelation. Self-report is participants' general description of their learning or rating behaviors and generally appears on questionnaires. Self-observation is participants' self-description of their actual instances of specific activities; thus, it is not as generalizable as the self-report data (Mackey & Gass, 2016). Self-revelation, also known as a *think-aloud*, is the participants' simultaneous description of their thought processes when they conduct the target activities.

Retrospective self-observation report is an appropriate method for the raters in the oral communication test. The main reason for using a self-observation report is that it can indirectly show the cognitive processes of raters when they evaluate test takers' performance. The self-revelation method may be more suitable for measuring the mental event in that it can be used before participants forget their cognitive processes; however, this method is not suitable for the raters in the speaking test, because it is not feasible for the raters to speak while listening to test takers' audio clips. Retrospective self-observation is the delayed inspection of what has happened; thus, there can be some biases due to the discrepancies between what raters have thought during the test time and what they think that they have thought during the testing situation.

CHAPTER 3. METHODOLOGY

This chapter describes the methodology of the present study. It begins with a description of the mixed-methods design for the current study, then describes the data source and the instruments used in the study and explains the procedures for data collection and data coding processes. The section concludes with a description of data analysis techniques that were used to answer the research questions.

3.1. Study Design

An explanatory sequential design with a mixed-methods approach was used in the current study. As shown in Figure *3.1*, quantitative data were collected and analyzed first, and then qualitative data was collected and analyzed to support the findings of the quantitative data analysis. For example, with regard to raters' rating scale use, a quantitative data analysis with many-facet Rasch measurement (MFRM) was used to analyze raters' score assignment depending on their understanding of task complexity; however, the quantitative analysis only verified the hypothesis regarding whether raters showed statistically different rating behaviors. While this statistical finding provided support for the hypothesis that raters' behavior difference was due to task complexity, a qualitative analysis for rater perceptions was needed to corroborate these findings. By integrating a qualitative analysis using retrospective verbal reports and interviews with raters, a more detailed explanation of rater behavior depending on task complexity was investigated. Figure *3.1* summarizes the mixed-methods design of the current study. More detailed information can be found in the following sections.

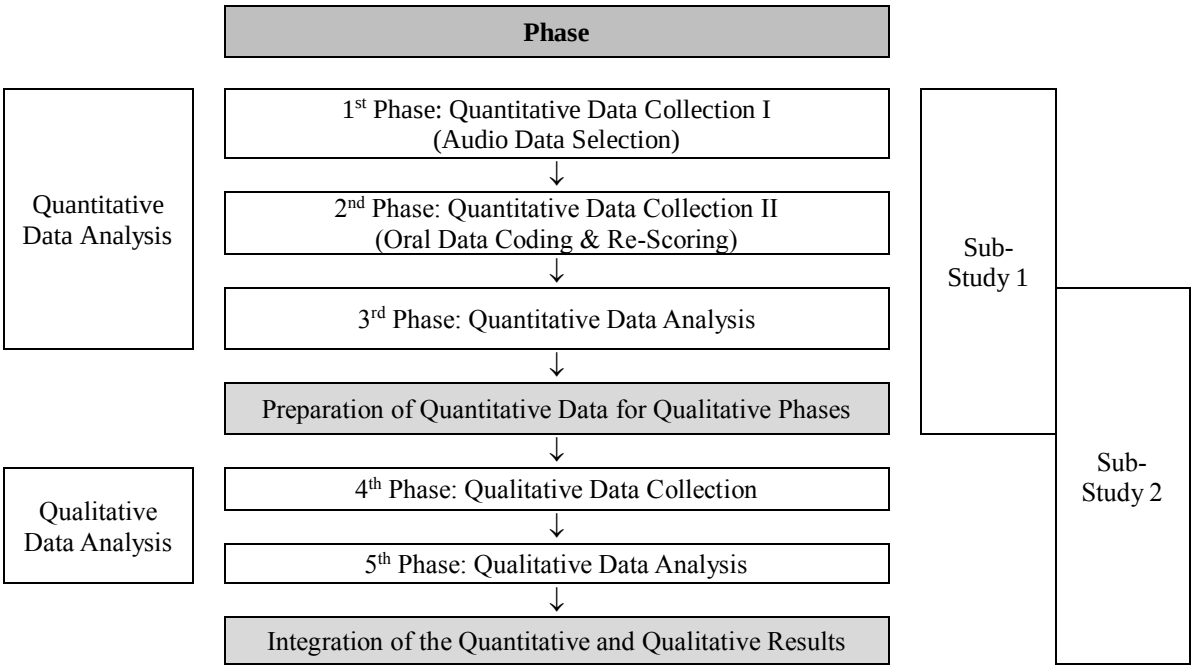


Figure 3.1. An exploratory sequential mixed-methods design of the current study (adapted from Creswell, 2014)

3.2. Data Source and Participants

Two types of data were used for Sub-Study 1: (a) operational rating data from 1,689 test takers whose native languages are not English and scored by 24 certified raters in the Oral English Certification Test (OECT), and (b) 162 audio recordings of 81 international graduate students in the OECT. Nine newly trained (novice) raters for this study participated in the current study and judged the 80 audio clips of 40 test takers for Sub-Study 2. Verbal reports and interviews with the nine novice raters were also analyzed.

3.2.1. Rating Data and Audio Clips

Operational OECT data

The extant operational Oral Proficiency Interview (OPI) rating data by the certified OECT raters and interviewers' task prompt selection logs were used for the current study.

The operational data contained a performance log of 1,689 test takers, 24 interviewers, and 24 experienced raters. The OECT is taken by international graduate students who meet the university's English requirement for admission, having achieved a score of 79 or above on an Internet-based Test of English as a Foreign Language (TOEFL iBT®) with a minimum score of 17 in both writing and speaking sections or an overall band score of 6.5 or above in the International English Language Testing System (IELTS) with 5.5 or above in all sub-sections. The students who took the ITA test did not meet requirements for the exemption criteria: scores of 27 or above on the TOEFL iBT listening and speaking sections or scores of nine on the IELTS listening and speaking sections. These score guidelines demonstrate that the test takers have acceptable levels of academic oral communication skills to study in U.S. universities, but are not assured their ability to teach in the classroom before being certified with the ITAs' oral communication test.

From the OPI and TEACH sections of the OECT, only the OPI section was used because this study investigated the relationship among tasks, interviewers, and raters and the TEACH section did not involve interviewers. As shown in Table 3.1, the OPI section consists of the following five tasks: a one-minute warm-up task, three two-minute impromptu interviews, and a two-minute role-play. The impromptu interview questions and the role-play topics were selected by the interviewer, who also served as one of the raters, based on his/her judgment of the test taker's proficiency level. In addition, the discussion topics and follow-up questions during the interview were selected by the interviewer; thus, it can be said that the materials were presented to test takers *adaptively* by interviewers. The adaptive selection of the prompt in the current study means that interviewers could select the interview prompts adaptively, depending on test takers' proficiency level within the given item pool.

Table 3.1. *Structure of the OPI*

Task	Phase	Purpose
	Warm-up (not scored)	Begin the interview
Impromptu speech 1	Level Check	Establish the floor
Impromptu speech 2	Probe	Establish the ceiling
Impromptu speech 3	Level Check or Probe	Establish the floor and/or the ceiling
Role-play	Probe and/or Level Check Transition (not scored)	Real life situation Close the interview and transition to TEACH simulation

(adapted from Cotos, 2014)

The impromptu speaking item sets for each participant consisted of three steps (level check, probe, and level check/probe) with three different topics, which consisted of five complexity levels with different linguistic target functions. The complexity level of the prompts followed the broad score bands in the scoring rubric (see Appendix A): *Level 1, Level 2, Level 3+, Level 3-,* and *Level 4.* Hence, interviewers were provided 15 question options to choose from in the impromptu speaking section (3 topic groups x 5 levels [*1, 2, 3+, 3-, 4*] = 15 questions). For the role-play part, interviewers had 12 situation options (3 situations x 4 levels [*1, 2, 3, 4*] = 12 situations). During the OPI session, interviewers were required to choose four out of 27 questions: three questions for the impromptu speaking and one topic for the role-play. For the current study, only the audio clips from speech produced in the impromptu speaking section were used, because the role-play section included a different test construct, one that invited more interviewer involvement in the task implementation. In addition, only *Level 1* to *Level 3+* prompts were used and coined as *high, mid,* and *low* task complexity, because *Level 3-* and *Level 4* prompts are rarely used in the OECT.

New rating data

Among the test takers in the operational OECT data from 2015 to 2018, test takers given three OPI prompts of different task complexity based on the original OECT task complexity guidelines were selected for new rating and linguistic analyses. One hundred twenty OPI prompts for 60 test takers were first selected and their task complexity was evaluated by three coders one a scale from 1-3 (1: *low*, 2: *mid*, and 3: *high*). The coders were non-native speakers of English; two of them held a linguistics-related college degree and one held a hard science degree. As multiple coders were involved in the coding, the inter-coder reliability coefficient with Krippendorff's alpha (Hayes & Krippendorff, 2007) was calculated to verify the agreement in coding. The *Krippendorff's alpha* for the task complexity coding was around .32, which must be regarded as low reliability (Krippendorff, 2004b). Thus, it can be argued that the agreements of the ratings for the task complexity coding were not satisfactory. To compensate for the weakness of task complexity coding by the three coders, the original task complexity coded by the OECT committee was compared to the average task complexity score of each prompt by the three coders. Out of 182 audio clips with 120 prompts, 80 audio clips from 40 test takers, based on the agreement between complexity levels by coders and by the OECT committee, were finally selected for further analysis. The final 80 audio clips were re-graded by the novice raters on a 0-13 holistic rating scale (see Attachment A) and by the experienced raters on a 0-13 analytic rating scale (fluency and lexico-grammar).

3.2.2. Spoken Data

The spoken data in this study were derived from audio recordings of the impromptu speaking tasks in the OPI section of the OECT. As was discussed in Section 3.2.1, 80 audio clips were selected for re-grading. The original length of each impromptu speaking session was two minutes, but for the current study, only the initial responses (about 60 seconds of the audio recordings) without any follow-up questions were used. This procedure was performed because it was determined that the first part of the test takers' response was more closely related to the prompt and less influenced by test takers' interaction with interviewers. The split audio files were reviewed and any sound that was not related to test takers' performance was muted, because any noise can reduce the performance of the Praat script (De Jong & Wempe, 2009) that automatizes the calculation of the linguistic measures. After cleaning the noise sounds from the audio file, the prompts read by interviewers and the responses by test takers were split to control the rating context. Raters were required to score test takers' proficiency by reviewing audio clips both *with* (*Question and Answer* rating context) and *without* listening to the prompts (*Answer Only* rating context).

3.2.3. Raters

The present study included two rater groups (novice and experienced). Nine novice raters (n = 9) were recruited and re-graded 80 audio clips of 40 test takers, after rater training, for further analysis with a holistic rating scale. Three experienced raters (n = 3), who went through the rater certification training, were also hired to grade the same 80 audio clips with an analytic rating scale.

Novice raters

Novice raters were all PhD students who were majoring in Applied Linguistics. They had each taken several language assessment courses and were trained as English Placement Test speaking section raters at a U.S. university, but had never worked or trained as an OECT rater prior to this study. They were also ESL instructors who had taught ESL courses for at least one year in a tertiary education institution in the U.S.

Experienced raters

Experienced raters were also all PhD students who were majoring in Applied Linguistics and had taught ESL courses for at least three years in a tertiary education institution in the U.S. Experienced raters had worked as OECT raters and had completed rater training prepared by the OECT committee. The rater training was a five-day (three hours per day) workshop consisting of an introduction to the OECT, practice rating of sample video clips, and live rating practice. The rater trainees had taken an introductory language assessment course and an English pronunciation teaching course. If the experienced raters were not native speakers of English, rater trainees had previously been scored at *Level 1*, the highest level, proficiency in the OECT. In the introductory session, the candidates were introduced to the OECT, including an orientation to assessment criteria, methods for conducting oral interviews, and test item development and evaluation. In the practice rating session, the trainees practiced rating using pre-recorded video clips and discussed their understanding of the evaluation criteria. In the live rating practice session, the trainees conducted live interviews and ratings in the same manner the certified raters did in a real testing situation. Once the trainees finished the five-day workshop, they were also required to

observe 42 test sessions, lasting approximately 30 minutes each, in preparation for taking a rater certification test. As the last step to becoming a certified rater, trainees evaluated ten video clips and reached an agreement of 70% or higher with the scores assigned by certified raters.

3.3. Instruments

The instruments for this study are the rating instruments for the OPI section of OECT, verbal report guidelines for raters, and written interview questions for raters.

3.3.1. Rating Instrument for the OPI

The scoring rubric (see Appendix A) of the OPI in the OECT includes four sub-categories (*functional competency*, *fluency*, *lexico-grammar*, and *pronunciation*), but the scores are given as a holistic score per each task (three impromptu speaking tasks and one role-play conversation task). The scoring rubric includes an 18-point scale, ranging from 13 (*lowest*) to 30 (*highest*) for each task with four cutoff scores for each level: 16 is the cutoff between *Level 4* and *Lower Level 3*; 18, between *Level 3*- and *Level 3*+; 20, between *Level 3*+ and *Level 2*; 22, *between Level 2* and *Level 1.* For the new rating data, the original scoring rubric on a 13-30 scale was converted to the new scoring rubric on a 0-13 scale to make sure that the new rating was less affected by raters' earlier exposure, if any, to the OECT.

The rating was conducted via Qualtrics software (Qualtrics, 2018) to control raters' access to task complexity of each audio clip during the rating session. On each page of the Qualtrics questionnaire, a single audio clip, rating rubric, and rating scale buttons (0-13) were displayed (see Appendix B). For the rating round with prompt and response audio clips,

raters were asked to click on the button to play each audio clip, listen to and evaluate the oral communication performance in the audio clip, and to click on and choose one grade on a 0-13 rating scale. For the rating round without prompt information, raters were asked only to play response audio clips before rating. If raters did not make decisions by clicking on the grade button, Qualtrics was set so as not to allow raters to move on to the next audio clip.

3.3.2. Verbal Report Guideline

Novice raters who re-scored the audio files with the holistic scoring scale were asked to report their cognitive processes during rating based on the retrospective verbal report guideline (see Appendix C for details). This verbal report guideline mostly focused on raters' score assignments depending on prompt complexity. Retrospective verbal reporting was conducted with the researcher within two weeks of their last rating session, so that raters would not forget about their general impression of the task's complexity as an oral communication rater.

3.3.3. Interview Questions for Raters

Raters were asked to answer the interview questions for each research question (see Appendix D for details). The interview questions measured how raters changed their score assignments depending on the complexity level of prompts and interviewers' behavior. The interview questions consisted of a questionnaire and interview. These interview questions supplemented the interpretation of the findings from the quantitative analysis.

3.4. Procedures

3.4.1. Overall Procedure

This study was conducted in five main phases, as illustrated in Figure *3.2*: (a) data collection, (b) oral data coding and re-scoring with the sample data, (c) quantitative data analysis, (d) rater interview and verbal reporting, and (e) qualitative data analysis and integration of quantitative and qualitative data analyses.

Phase	Procedure	Product
1st Phase: Quantitative Data Collection I (Rating Data Selection) (see Section 3.4.2 for details)	• Operational data selection from OECT database	• Numeric data
↓		
2nd Phase: Quantitative Data Collection II (Oral Data Coding & Re-scoring) (see Section 3.4.3 for details)	• Coding of interviewers' prompt selection • Rater training • Scoring of the split audio files (about 60 seconds long)	• Numeric data (Split audio files, score data)
↓		
3rd Phase: Quantitative Data Analysis (see Section 3.4.4. for details)	• Data screening • Logistic regression analysis • Paired samples *t*-test or Wilcoxson rank-sum test • Many-facet Rasch measurement analysis	• Descriptive statistics • Regression coefficient • Variance component • Rating scale threshold and item (or task) difficulty
↓		
Prepare Quantitative Data for Qualitative Phases	• Selection of audio clips for each rater for retrospective verbal report	
↓		
4th Phase: Qualitative Data Collection (see Section 3.4.5)	• Interview of raters • Retrospective verbal report using interview audio clips	• Interview transcript • Questionnaire report
↓		
5th Phase: Qualitative Data Analysis (see Section 3.4.5)	• Retrospective verbal reports and interview analysis	• Cognitive process report for the quantitative analysis results
↓		
Integration of the Quantitative and Qualitative Results	• Interpretation of both quantitative and qualitative results	

Figure 3.2. An illustration of the dissertation study process

In the data collection phase, the extant audio files of the OECT rated with a holistic score were chosen. Audio files from the OECT database were selected as sample data and divided into smaller segments for analysis. In the oral data coding and re-scoring phase, the order of the split audio files was randomized before being re-scored by the raters to minimize the contrast effects between preceding audio clips and the audio clips currently being evaluated. Contrast effects refer to "the influences of previous stimuli on the evaluation or judgment of a new stimulus" (Daly & Dickson-Markman, 1982, p. 309). In the rater interview and verbal reporting phase, raters were asked to answer questions about their perceptions of the interview prompts. In the last phase of the data analysis, the scores and the coding data were quantitatively and qualitatively analyzed.

3.4.2. 1st Phase: Quantitative Data Collection

The first phase of the study involved a selection of the sample data from the extant audio files of the OECT and operational rating data for coding of prompt complexity.

Audio data selection

The audio files of samples rated with a holistic score on the scoring rubric (see Appendix A) by human raters were collected from the Oral Proficiency Interview (OPI) section of the OECT at Iowa State University using the following steps. Approximately 1,689 test takers' data collected between spring 2016 and spring 2018 were selected for this study. For further analysis, the scoring data and results were reviewed by the researcher to verify the data integrity according to the following criteria. First, test takers' data that were not scored by three raters were excluded. Second, test takers who were not given prompts of

multiple task complexities (e.g., *high* and *low*) were excluded. Finally, audio clips with a poor recording quality also were excluded. Among the 1,689 test takers' data recordings, approximately 162 audio clips of 81 test takers (81 test takers x 2 tasks = 162 audio clips), who were given prompts of multiple complexity levels, were selected for further data coding. The operational data of 1,689 test takers were reviewed by the researcher to analyze interviewers' prompt use, and 162 audio clips were used for linguistic feature analysis and for the new rating.

Split the sample files into smaller pieces with initial and follow-up questions

The duration of each audio file was approximately two minutes and started with interviewer's prompts for the impromptu speaking tasks. The original file was divided into smaller audio clips by interviewers' follow-up questions. For example, if the interviewer asked a follow-up question during the two-minute test time, the audio clip was cut right before the interviewer's follow-up question. Audio files with follow-up questions were not used for the current study, because follow-up questions were not consistent across prompts and raters. Only the initial response files (approximately 60 seconds long), which were directly affected by interviewers' prompt selection, were used. Each audio clip then was re-scored in a later phase to investigate the impact of question types on test scores.

3.4.3. 2nd Phase: Oral Data Coding and Re-scoring

Interviewer's prompt coding

The researcher listened to the 162 audio files again and coded the information on the prompts that interviewers used during the impromptu speaking session according to

complexity level (i.e., *high*, *mid*, and *low*) and question type (e.g., narration, description, hypothesis) of the prompts. As described in Section 3.2.1, the coding was initially conducted by three coders. Due to the lack of inter-coder reliability in the task complexity coding, the final complexity level of each prompt was coded by comparing the average coding score by three coders with the complexity level provided by the OECT committee. If the two levels provided by three coders and the OECT committee were same, the agreed level was used as the final level of the prompt. If the levels by the three coders and the OECT committee did not agree, the final decision was made by averaging the average coding scores by the three coders and the complexity levels by the OECT committee.

Rater training for the holistic scoring

The nine newly recruited (novice) raters participated in rater training approximately two days prior to the online rating session following the following format. First, the researcher explained the construct of the OECT, its format, and its scoring rubric for 10 minutes. Second, raters were given 22 audio clips whose original score in the OECT ranged from 3 to 13 in the transformed rubric on a 0-13 holistic rating scale (see Appendix A). Raters then participated in a calibration session with other trainees to compare their individual decisions. For each audio clip, one rater initiated the conversation and shared his or her score, and other raters compared their scores and had a short discussion about the sample and the ratings. During the discussion, however, raters were not guided to disclose the rationale behind their rating decision, because this opinion sharing could have spoiled the goal of this study, which focuses on the effect of raters' understanding of task complexity on their scoring. Finally, raters were given the same training audio clips again online to review

before starting holistic grading of 80 audio clips. In total, the rating session lasted approximately one hour and twenty minutes.

For the analytic scoring, three experienced raters underwent a shortened version of rater training, which only lasted about 30 minutes, because they were certified OECT raters and were already familiar with the test. As in the training session with a holistic scale, raters were given 22 audio clips to share their scoring. As opposed to the holistic rating session, raters in the training with an analytic scale were given an opportunity to share their rationale for their score decision, because analytic scoring was not used to examine the effect of task complexity on scoring. Raters for the analytic scoring were asked to focus on each analytic scoring criterion (*lexico-grammar* and *fluency*) and not to be influenced by holistic and/or another analytic scoring criterion.

Transcription and coding of linguistic features

The selected 162 audio recordings of test takers' performance on the OECT were transcribed and coded for statistical analysis. Any inaudible words during the transcription process were not transcribed, but instead coded with the number of syllables, based on the vowel sound, to be used for further analysis. The Analysis of Speech unit (AS-unit), a syntactic unit consisting of an independent clause with any subordinated clauses (Foster, Tonkyn, & Wigglesworth, 2000), was used as the smallest unit of analysis, because it is more suitable than the T-unit or c-unit when it comes to spoken data analysis (Foster et al., 2000; Norris & Ortega, 2009; Plough, Briggs, & Van Bonn, 2010). The clauses and AS-units were compared to the output of the automatic syntactic analysis with a Stanford CoreNLP Toolkit (Manning et al., 2014) to verify and support the hand coding by the researcher. To understand

the relationship between proficiency scores graded by human raters and linguistic features in the spoken data depending on the task complexity level, the speech data were linguistically analyzed with two analytic scoring criteria: *lexico-grammar* and *fluency*. *Pronunciation* scoring was not linguistically analyzed, because task complexity was not assumed to affect the quality of pronunciation.

Re-scoring with the split audio files

Among the 162 audio clips that were transcribed and linguistically analyzed, 80 audio clips were selected for the analysis of rating scale use. Raters were asked to score the 80 selected audio clips (approximately 60 seconds; 40 test takers x 2 tasks = 80 audio clips) of performances based on prompts of different complexity levels via the online survey platform Qualtrics. The novice raters (n = 9) were first asked to score audio clips based on the transformed 0-13 holistic scale (see Appendix A). In the first round of rating, raters listened to the test takers' responses only without having access to the task prompt. In the second round, raters were asked to re-score the audio clips with task prompts, which would have helped raters to understand how complex the original task prompt was. The second round of rating was conducted at least five days after the initial scoring to reduce any errors related to the effect of the first rating on the raters' second rating.

For the analytic scoring, experienced raters (n = 3) were asked to score the audio data using a 0-13 analytic scale. The analytic scoring was also completed by rating one analytic scale category per week to reduce the halo effect among the scores with different rating scales. Halo effect refers to "raters' tendency to assign ratees similar ratings on conceptually distinct traits" (Myford & Wolfe, 2004, p. 209). For analytic scoring, raters were asked to

conduct scoring in the order of *fluency* and *lexico-grammar* scoring sessions. *Functional competency* and *pronunciation* were not used for the current study, even though they were included in the evaluation rubric, because *pronunciation* would not change depending on task complexity and a 60-second audio clip would not be sufficient for raters to evaluate the sample's content development.

The audio clips were mixed together and the order of scoring was randomized to minimize an effect because of the order of audio file rating (e.g., *halo effect*, *contrast error*, *proximity error*; Saal, Downey, & Lahey, 1980; Van Moere, 2013). A counter-balanced design following a Latin square approach (Box, Hunter, & Hunter, 2005; Tabachnick & Fidell, 2013) with blocks of 20 audio clips was employed to randomize the rating order.

3.4.4. 3rd Phases: Quantitative Data Analysis

Before eliciting raters' verbal reports and conducting the interviews, the data were analyzed with quantitative data analysis methods, such as multilevel ordinal logistic regression, paired samples *t*-test, Wilcoxon signed-rank test, and MFRM. Detailed quantitative data analyses are described in the data analysis section (see Section 3.5).

3.4.5. 4th & 5th Phase: Qualitative Data Collection and Analysis

Retrospective verbal report

After the re-scoring session with 80 audio clips, the nine novice raters were asked to participate in the verbal reporting session, consisting of about 10-minute training and 50-minute think-aloud session. In the verbal reporting session, using a retrospective verbal reporting method, raters reviewed eight selected audio clips and they were asked to describe

their cognitive processes while scoring as they listened to the audio clips. Raters first listened to the full 60-second audio clip and judged its proficiency score based on the holistic rating rubric. Raters then listened to the same audio clip again, but the researcher stopped the audio clip every 20 seconds and asked raters to describe their rationale behind their proficiency scoring. For example, raters described how they felt when test takers did not communicate effectively (e.g., stumbling, hesitating, or using incorrect vocabulary). The verbal report data were qualitatively analyzed by focusing on raters' perception of the effect of task complexity on their scoring. The detailed procedure of retrospective verbal reporting is presented in Appendix C.

Interview questions

After the verbal reporting session, raters were asked to answer how they applied the rating scales in their rating and their perceptions of the rating scales. Raters were also asked to answer the interview questions on their evaluation of the interviewer's behaviors in relation to task complexity, such as *topic choice, poor topic development, closed questions,* or *speech style* (A. Brown, 2005), which might have influenced their rating behavior. The interview data were qualitatively analyzed in relation to each rater's rating behavior discovered in the quantitative data analysis. The interview questions are presented in Appendix D.

3.5. Data Analysis

3.5.1. Interviewers' Adaptiveness in the OECT

In order to understand how raters adjust their rating severity depending on the complexity level of the prompt in the OECT, an adaptive performance-based oral communication test, the adaptiveness of the test was first investigated. The adaptiveness of the test was verified by investigating how interviewers considered the scores of the preceding tasks when they decided which complexity level to apply in the following tasks.

As the selection of task complexity level (*high*, *mid*, and *low*), which is ordinal, was used as a dependent variable and the score graded by an interviewer in the preceding task, which is continuous, was used as an independent variable, an ordinal logistic regression model was chosen instead of linear regression or simple logistic regression (Agresti, 2002; O'Connell, 2006). In addition, as interviewers were independent of each other in terms of their ways for selecting prompts with different complexities, the dependent variable was nested within the interviewers. Thus, the interviewer variable was set as a group variable and the two-level ordered category response model, or multilevel ordinal logistic regression (O'Connell, 2006), was finally adopted to test the adaptiveness of the OECT.

The multilevel ordinal logistic regression model identified the conditions in which interviewers were motivated to ask *higher*, *equal*, or *lower* complexity level questions during the impromptu speaking session. The ordinal regression model used in the current study, the random cumulative logit model, can be expressed as follows:

$$\text{Log}[\frac{\text{P}(y_{ij} \le k)}{\text{P}(y_{ij} > k)}] = \text{logit}(\gamma_{kij}) = \alpha_k + \beta\text{SCORE}_{ij} + u_{0j} + u_{1j}\text{SCORE}_{ij}\ ,\ k = 1, 2 \qquad (3.1)$$

$$\begin{pmatrix} u0j \\ u1j \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} \sigma\text{^}2_{u0} & \\ \sigma_{u01} & \sigma\text{^}2_{u1} \end{pmatrix}\right\}$$

where

$\text{P}(y_{ij} \le k)$ = probability that the observation *i*, or complexity level, is at or less than category *k*

α_k = threshold parameters

β = coefficient of SCORE variable

SCORE = score (13-30) of the preceding task graded by an interviewer

u_{0j} = interviewer intercept effect

u_{1j} = interviewer slope random effect

The interviewer intercept and slope random effects (u_{0j} and u_{1j}) are assumed to bivariate normally distributed with zero means, variance σ^2_{u0} and σ^2_{u1}, and covariance σ^2_{u01}. The slope of the linear relationship between SCORE and the log-odds that $y \le k$ is $\beta + u_{1j}$. The model was fitted by allowing the effect of explanatory variable SCORE to vary across all interviewers.

Prior to the analysis, the dependent variable was assessed to have an ordinal data type as discussed in Section 3.2.1. Multicollinearity of the independent variables was also assessed, but it was not deemed a concern, because the final model was set as the model with a single independent variable. Finally, the proportional odds assumption with an explanatory variable of ordinal regression was examined to check whether ordinal regression could be used with the interviewer data in the OECT. The details of the assumption test for ordinal regression model are discussed in Section 4.1.1.

3.5.2. Effect of Task Complexity

Wilcoxon signed-rank test (Wilcoxon, 1945) and a paired-samples *t*-test were employed to examine the effect of task complexity on the oral task performance in terms of linguistic complexity and fluency and on two categories of proficiency scores (i.e., *lexico-grammar*, *fluency*). Multiple univariate analyses (e.g., Wilcoxon signed-rank test, paired-samples *t*-test), instead of multivariate analysis (e.g., MANOVA), were used, because this study focused on the effects of task complexity on each dependent variable (i.e., complexity and fluency measures) independent of the other dependent variables (Huberty & Morris, 1989).

Task complexity with two levels (*high* and *low*) was used as an independent variable, and the linguistic outputs or proficiency scores were used as a dependent variable. For the analysis of the effect of task complexity on linguistic features, a Wilcoxon signed-rank test, a non-parametric equivalent of paired samples *t*-test, was employed, because the distribution of linguistic features did not meet the assumption of normality in parametric tests. Prior to the analysis, the distribution of the linguistic feature frequency difference between *high* and *low* complexity prompts was assessed to check the assumption of the Wilcoxon signed-rank test. For the analysis of the effect of task complexity on proficiency scores, a paired-samples t-test was employed because the proficiency scoring data generally met the assumption of the parametric test. The details of the assumption test for Wilcoxon signed-rank test and paired-samples t-test are discussed in Section 4.2.1 and Section 4.2.2.

3.5.3. Variation of Rating Severity by Task Complexity

The purpose of using a task-complexity-related three-facet partial credit model (PCM) MFRM is to understand how differently raters used the rating scale depending on the task complexity. The partial credit model used in the current study can be expressed as follows:

$$\mathrm{Ln}\left[\frac{P_{nljk}}{P_{nljk-1}}\right] = \theta_n - \delta_l - \alpha_j - \gamma_c - \varphi_{lc} - \tau_{lk}, \quad (3.2)$$

where

P_{nljk} = probability of test taker n receiving a rating of k from rater j for task l,

P_{nljk-1} = probability of test taker n receiving a rating of $k-1$ from rater j for task l,

θ_n = speaking ability of test taker n,

δ_l = difficulty of the task complexity l,

α_j = severity of rater j,

γ_c = rating context c

(Dummy variable; *Answer Only* and *Question and Answer* contexts)

φ_{lc} = interaction between task complexity l and rating context c,

τ_{lk} = difficulty of scale category k relative to scale category $k-1$ on task complexity l.

The partial credit model in Equation 3.2 provided the structure of the rating scales for each task complexity level and helped compare how raters appropriately applied their rating scales coupled with task complexity. The task complexity-specific category threshold estimates (i.e., Rasch-Andrich thresholds; Bond & Fox, 2015; Linacre, 2014) and their standard errors indicate how closely or differently the rating scales were used by raters when the task complexity is different. For example, when the locations of category threshold estimate of *high* complexity tasks are statistically higher than those of *low* complexity tasks, test takers would receive higher scores when they were administered *high* complexity tasks than *low* complexity ones.

Prior to data analysis with the MFRM model, the assumptions of local independence of items, unidimensionality, the absence of guessing, and equal item discrimination were assessed (Bond & Fox, 2015). First, Rasch-Cohen's Kappa (Linacre, 2014) was used to test the independence of the data. Rater training and randomization of the rating order were also considered to confirm the independence of the data. Second, the unidimensionality assumption was tested by examining the infit and outfit statistics for task complexity with an acceptable range of 0.5 to 1.5 (Linacre, 2014). Finally, the assumptions of the absence of guessing and equal item discrimination were assumed to be met in this study, because rater training was vigorous and raters generally were trained not to guess the scores in the test. Thus, only the details of the local independence and unidimensionality assumptions are discussed in Section 4.1.2 and Section 4.3.1.

For the current study, two sets of MFRM models were employed to examine raters' rating scale use in an adaptive performance-based oral communication test: one with the operational data from the OECT and the other with the selected 80 audio files in an experimental context. The first MFRM model was modeled to see how raters behaved in the real adaptive oral communication test context by using operational OECT data. As was introduced in Section 3.2.1, test takers in the OECT were given, in a row, three tasks of diverse task complexity. Because of this consecutive rating for a test taker with multiple items, the assumption of local independence may be violated in the operational OECT data. To overcome the potential violation of the independence assumption, the scoring data with the first task of the three consecutive prompts was modeled to assess the rating scale use by raters. After the data analysis with the first task, the same model with the full data, which

include the second and the third tasks, were also modeled to test if the local independence assumption was held in the OECT data.

Another MFRM model with the data in an experimental context was analyzed to examine raters' rating scale use depending on the task complexity coupled with their understanding of task complexity. Two separate MFRM analyses for two types of raters were conducted: one for the raters who responded in an interview that they considered task complexity in their rating and another for those who indicated they did not consider task complexity. The details of the experimental context were introduced in Session 3.4.3. As was analyzed with the operational OECT data, the Rasch-Andrich thresholds of the new MFRM model with the experimental context data were used to identify how closely or differently raters used the rating scale depending on the task complexity. The task complexity-specific category threshold estimates and standard error were used to measure how closely the rating scales were used by raters when the task complexity was different. Welch's *t*-tests (Welch, 1947) were also conducted to identify any interaction between task complexity (*high*, *equal*, and *low*) and rating context (rating *with* or *without* knowledge of task complexity). The rating context was used as the independent variable and the difficulty score of each complexity level was used as the dependent variable. The pairwise bias report presented the difficulty estimate of each complexity prompt when evaluated in the *Answer Only* context or *Question and Answer* rating context, the difference between the difficulties of each complexity prompt, and an effect size of the difference. For example, if the difficulty score of *high* complexity in the rating context *with* the knowledge of task complexity (*Answer Only* rating context) is statistically higher than that in the rating context *without* the knowledge of task complexity (*Question and Answer* rating context), it can be claimed that the average task difficulty of

high complexity prompts decreased from the *Answer Only* to *Question and Answer* rating contexts.

3.5.4. Data Analysis Software

Two sets of software were used to process audio files, and three statistical packages were used to analyze the data. First, to split the mp3 files from the impromptu speaking in the OPI, Direct WAV MP3 Splitter (Piston Software, 2017) was used. This splitter software automatically detects silence in the audio file, thus helping save time for researchers in splitting the audio files. Second, to calculate temporal measures (e.g., *Speech Rate, Phonation-Time Ratio*) of the audio files, a Praat script (De Jong & Wempe, 2009) was used, as the Praat script can automatize the calculation of the linguistic measures. Third, to conduct the paired samples *t*-test and its non-parametric equivalent test, the Wilcoxon signed-rank test, for the analytic score and linguistic feature comparison, IBM® SPSS® 22 (Arbuckle, 2013) was used, because SPSS is one of the most widely used statistical packages for group mean comparison in the social science studies. Fourth, to investigate the adaptiveness of interviewers' selection of prompt complexity, MLwiN version 2.36 (Rasbash, Browne, Healy, & Cameron, 2015) was used, because MLwiN can easily analyze both single and multi-level ordinal regression and can produce various figures for the data analysis. Finally, to analyze the MFRM models in the study, FACETS version 3.80 (Linacre, 2014) was used, because FACETS is the most accessible program for analyzing rating scale data while taking into account judge mediation (McNamara, 1996).

CHAPTER 4. RESULTS

This chapter discusses the results of the data analysis conducted to answer the research questions using the findings from both quantitative and qualitative analyses described in the previous chapter. Sections 4.1 and 4.2 reports on the findings of the operational data analysis, and Sections 4.3 and 4.4 report the results of the experimental data analysis. This chapter begins with the characteristics of adaptiveness in the Oral English Certification Test (OECT) rating procedure by using ordinal regression (see Section 4.1). It then investigates experienced raters' use of the rating scale in the OECT based on many-facet Rasch measurement (MFRM) (see Section 4.1) and test takers' linguistic performance depending on the task complexity based on paired samples *t*-tests (or the Wilcoxon signed-rank test) (see Section 4.2). Finally, it describes newly trained raters' use of rating scales depending on their knowledge of task complexity based on the analysis of MFRM (see Section 4.3) and verbal reports (see Section 4.4).

4.1. Rating Severity Variation in an Adaptive Performance-Based Oral Communication Test (Sub-Study 1)

This section reports on the analysis targeted at answering the first research question "RQ 1: How do raters adjust their score assignment depending on the complexity level of the prompts in an adaptive performance-based oral communication test?" To answer this question, the adaptiveness of the oral communication test, the OECT, was first examined by using logistic regression models for ordinal responses, then variation of the raters' rating scale use was investigated using MFRM.

4.1.1. Interviewer's Adaptive Selection of Test Prompts

As this study investigates the rating severity variation in an adaptive oral communication context, the adaptiveness of the test prompt was first studied by examining the interviewer's selection of tasks of different task complexity levels in the operational OECT scoring data.

Descriptive statistics for OECT scores by task complexity

The descriptive statistics for the oral communication test scores by task complexity are presented in Table 4.1. The mean score of test takers who were given *high* difficulty prompts in the following task was 23.28 on a 13-30 holistic scale, a value higher than *mid* and *low* by 2.24 and 4.03, respectively.

Table 4.1. *Descriptive Statistics for OECT SCOREs by Task Complexity Level (N=1,689)*

Task Complexity Level	Number of Ratings (%)	Mean	Standard Deviation
high	617 (22.0%)	23.28	1.57
mid	700 (41.4%)	21.04	1.62
low	372 (36.5%)	19.25	1.42
Total	1,689 (100%)	21.46	2.19

Multilevel models for ordinal responses

A multilevel logistic regression model for ordinal responses (Agresti, 2007; O'Connell, 2006) was fitted to the data to examine the adaptiveness of the test. The model was fitted using the Markov chain Monte Carlo (MCMC) method in MLwiN (Rasbash et al., 2015), as the method produces less biased results when the cluster size, or the number of students interviewed by each interviewer, in the data is small (Browne, 2015). The performance score, a level-1 independent variable, was centered around the mean in each

interviewer (as a group), because group-mean centering is more useful for examining the association between a level-1 predictor and the outcome variable (Barkaoui, 2013; Enders & Tofighi, 2007). Table 4.2 summarizes three models compared in the current study.

Assumptions of multilevel ordinal regression. All assumptions were satisfied for the multilevel logistic regression for ordinal responses. First, the dependent variable was measured at the ordinal level (*low*, *mid*, and *high* levels). Second, there were no multicollinearity issues, because there was only one independent variable without any level-2 independent variables. Finally, the proportional odds assumption, with an explanatory variable, of ordinal regression was shown to be tenable, as described in Model 2.

Table 4.2. *Results for Three Multilevel Ordinal Models (Cumulative Odds)*

	Model 1		**Model 2**		**Model 3**	
	Coeff (SE)	**Prob.**	**Coeff (SE)**	**Prob.**	**Coeff (SE)**	**Prob.**
Fixed Effects						
Intercept (α_1) (=low)	-1.45 (0.15)	0.19**	-2.61 (0.24)	0.07**	-2.71 (0.31)	0.06**
Threshold (α_2) (≤mid)	0.48 (0.15)	0.62**	0.78 (0.23)	0.69**	0.78 (0.29)	0.68**
SCORE (β_1)			-1.04 (0.04)	0.26**	-1.15 (0.11)	0.24**
Random Effects						
Var. in Intercepts (σ^2_{u0})	0.37 *		1.29*		1.71*	
Cov. Intercepts * Slopes (σ_{u01})					0.14	
Var. in SCORE Slopes (σ^2_{u1})					0.16*	
DIC (*pD*)	**3504.14 (17.94)**		**2381.63(22.26)**		**2331.75 (35.97)**	

Notes. MCMC estimation (iteration = 10,000); group-mean centering of SCORE; *p < .05, **p < .01; DIC = Deviance Information Criterion; *pD* = the effective number of parameters

Model 1 (Null two-level model). First, an intercept-only model in the Equation 4.1, a *null model* with no predictors, was estimated as a benchmark when comparing with other models (Hox, 2010; Leckie & Baird, 2011):

$$\text{Log}[\frac{\Pr(y_{ij} \le k)}{\Pr(y_{ij} > k)}] = \text{logit}(\gamma_{kij}) = \alpha_k + u_j, \qquad k = 1, 2 \qquad (4.1)$$

$$u_j \sim N(0, \sigma^2_u)$$

where u_j is the normally distributed interviewer random effect with a mean of zero and variance σ^2_u; y_{ij} is the ordinal response for a test taker *i* interviewed by an interviewer *j*; α_k represents the threshold parameters which are interpreted as the log-odds that a test taker *i* has a response of *k* or lower (1 = *low*, 2 = *mid*) with no variance of interviewers ($u = 0$). As shown in Table 4.2, the between-interviewer variance in the null model was estimated as 0.37, which implies a *intraclass correlation coefficient* (ICC) of 0.101. ICC was calculated as the ratio of interviewer variance to the total variance: $\frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e} = \frac{0.37}{0.37 + 3.29} = 0.101$, where $\sigma^2_e = 3.29$ for the standard logit model ($\pi^2/3 = 3.29$). Thus, approximately 10.1% of the variance in the task complexity selection was due to between-interviewer variation. Even though between-interviewer variation was not large, multilevel modeling was maintained, because the data used for the current study had a multilevel structure (Nezlek, 2008).

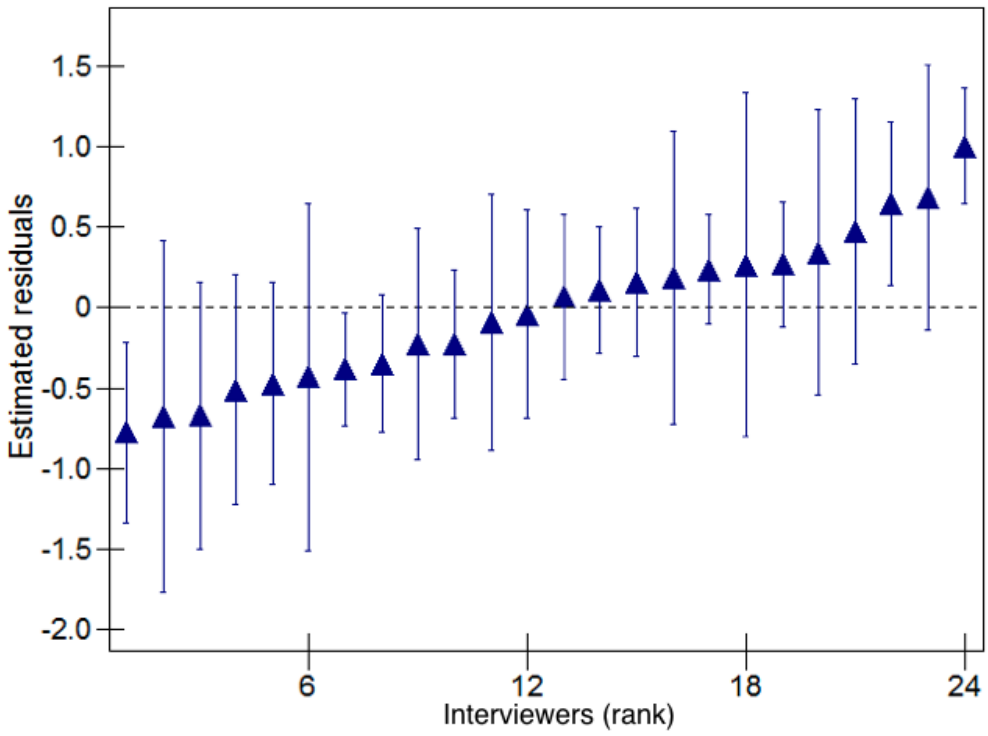


Figure 4.1. Estimated residuals for the 24 interviewers in the OECT

The cumulative logit prediction on average across interviewers for $y_{ij} \leq 1$ (= *low*) was -1.45 and $y_{ij} \leq 2$ (= *mid*) was 0.48. The interviewer random effects u_j allows the cumulative response probabilities to vary across interviewers. Figure *4.1* shows the estimated residuals for the 24 interviewers in the OECT data from 2015 to 2018. The 95% confidence interval of some interviewers does not overlap the horizontal line at zero, indicating that task complexity level in the OECT is significantly above or below average at the 5% level for these interviewers.

Model 2 (Random intercept cumulative logit model). Model 2 was fitted by including an additional test-taker level explanatory predictor SCORE, a test taker *i*'s performance score judged by an interviewer *j* in the preceding task:

$$\text{Log}[\frac{\Pr(y_{ij} \leq k)}{\Pr(y_{ij} > k)}] = \text{logit}(\gamma_{kij}) = \alpha_k + \beta \text{SCORE}_{ij} + u_j, \qquad k = 1, 2 \qquad (4.2)$$

$$u_j \sim N(0, \sigma^2_u)$$

where u_j is a normally distributed interviewer random effect with a mean of zero and variance σ^2_u. As shown in Table 4.2, the cumulative logit prediction on average across interviewers for $y_{ij} \leq 1$ (= *low*) was -2.61 and $y_{ij} \leq 2$ (= *mid*) was 0.78. The parameter β indicates that the effect of a one-unit increase of SCORE on the log-odds that $y \leq k$ while controlling the effect of interviewers (u) was -1.04.

Testing the proportional odds assumption of SCORE. The proportional odds assumption of ordinal regression implies that the effect of SCORE on the log-odds of being in the task complexity level *k* or below is constant across different levels *k.* A Wald test was conducted to test the proportional odds assumption of the SCORE effect by allowing the coefficient of SCORE to vary according to the response category *k*. The null hypothesis of a test that the effects of SCORE are proportional is H_0: $\beta_1 = \beta_2$ in the following Equation 4.3:

$$\text{logit}(\gamma_k) = \alpha_k + \beta_k \text{SCORE}, \qquad k = 1, 2 \qquad (4.3)$$

$H_0: \beta_1 = \beta_2$

The Wald test failed to reject the null hypothesis that the effects of SCORE are the same for all response categories ($\chi^2_{(1)} = 1.99$, $p = .16$). Thus, the proportional odds assumption holds and the SCORE variable is added to the model with a common coefficient.

Model 3 (Random slope cumulative logit model). As in Equation 4.4, Model 3 was fitted by allowing the effect of explanatory variable SCORE to vary across all interviewers:

$$\text{Ln}\left[\frac{\Pr(y_{ij} \le k)}{\Pr(y_{ij} > k)}\right] = \text{logit}(\gamma_{kij}) = \alpha_k + \beta \text{SCORE}_{ij} + u_{0j} + u_{1j}\text{SCORE}_{ij}, \; k = 1, 2 \qquad (4.4)$$

$$\begin{pmatrix} u0j \\ u1j \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma\text{^}2_{u0} & \\ \sigma_{u01} & \sigma\text{^}2_{u1} \end{pmatrix}\right\}$$

where u_{0j} and u_{1j} denote the interviewer intercept and slope random effects, and they are assumed to be bivariate and normally distributed with zero means variances σ^2_{u0} and σ^2_{u1} and covariance σ_{u01}. In Model 3, the slope of the linear relationship between SCORE and the log-odds that $y \le k$ is $\beta + u_{1j}$. The quadratic term of SCORE was initially modeled in Equation 4.4, but it was deleted because the quadratic term was not conceptually relevant to task complexity selection and its contribution to the model was not statistically significant. Even though a Wald test statistic for the effect of SCORE on the selection of task complexity level across interviewers was not statistically significant ($\chi^2_{(2)} = 4.14$, $p = .13$), the SCORE variable was retained because adding the variable as a random slope significantly improves the model fit of Model 3 from Model 2 ($\text{DIC}_{\text{Model 2}} - \text{DIC}_{\text{Model 3}} = 49.88$). Thus, Model 3 was chosen as the final model to answer RQ 1-1: To what extent does the performance (in holistic scores) of test takers in each task affect interviewers' selection of the complexity level of the following tasks in an adaptive performance-based oral communication test?

The results for Model 3 show that the slope of the linear relationship between SCORE and the log-odds that $y \leq k$ was -1.15 + u_{1j} for interviewer *j*. The between-interviewer variance in the effect of SCORE was estimated as 0.16 ($\chi^2_{(1)} = 4.139$, p = .042). The intercept variance was estimated as 1.71 ($\chi^2_{(1)} = 6.227$, p = .013), which is the between-interviewer variance at the interviewer group-mean. The intercept-slope covariance estimate of 0.14 was not statistically significant, which means that there were no correlations between the interviewers with above average task complexity (intercept residual) and an average effect of SCORE (slope residual).

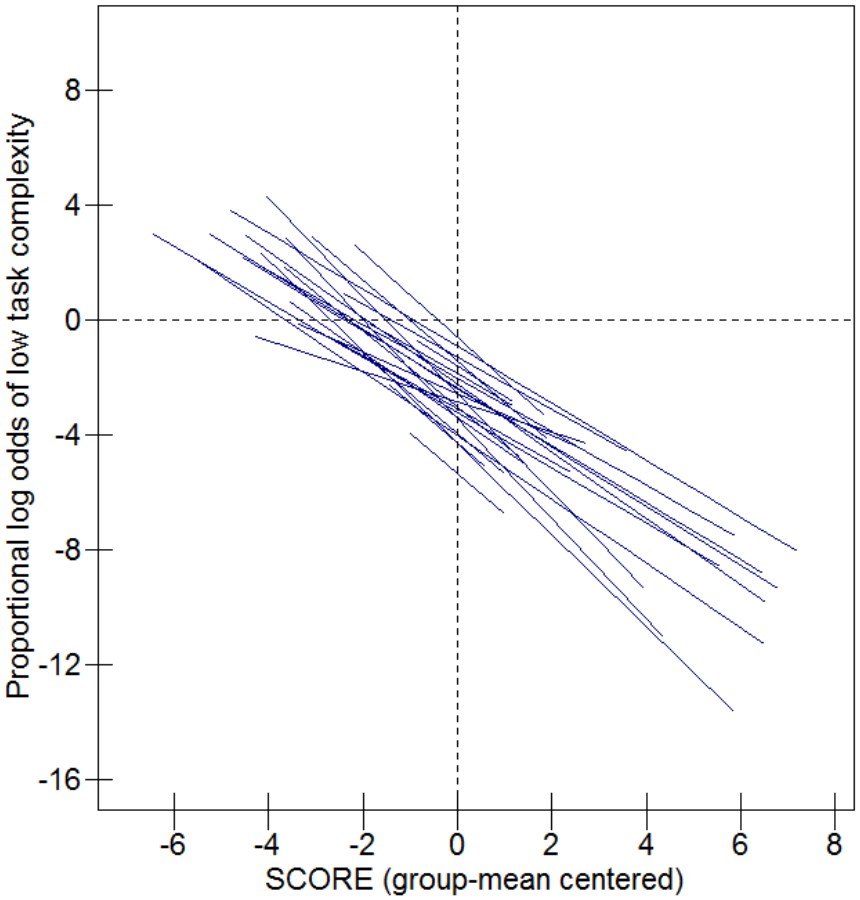


Figure 4.2. Effect of SCORE on the log-odds of task complexity level in the low category

Figure *4.2* shows the effect of SCORE on the log-odds of interviewers' selection of task complexity being in the low category. As the SCORE effect is assumed to be proportional, as shown in Model 2, the log-odds of being in the low category also shows how the SCORE affected the selection of task complexity in general by each interviewer. As shown in Figure *4.2*, the predicted curves were more spread out vertically at both ends of the scale (*high* and *low*) than around average SCORE, which implies that variability in the log-odds of having a low category in the OECT increases as SCOREs are away from the mean.

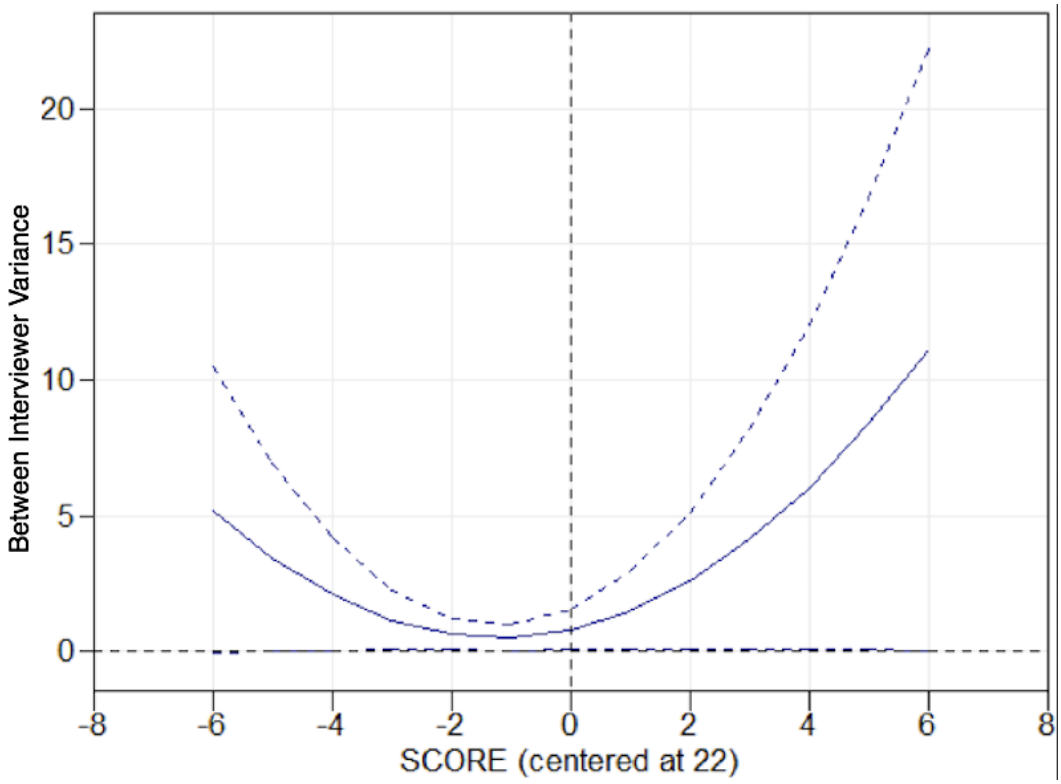


Figure 4.3. Between-interviewer variance of task complexity by SCORE

To better understand the variability of each interviewer's selection of task complexity in comparison to students' performance score, Model 3 was re-fitted with the SCORE variable centered at 22, which is the threshold from Level 2 proficiency (medium level in the OECT) to Level 1 proficiency (highest level in the OECT). Figure *4.3* shows that the

between-interviewer standard error increases or decreases as a function of SCORE when the SCORE reaches both ends. The dotted lines indicate a 95% confidence interval of interviewer variance. The wider standard error in the test score may be due to the small number of test takers at each end of the score range. As the cut-off scores for *mid* and *high* levels in the OECT are 19 and 22, the higher standard error at each end of score range would not be a serious issue for explaining an interviewer's selection of task complexity in general.

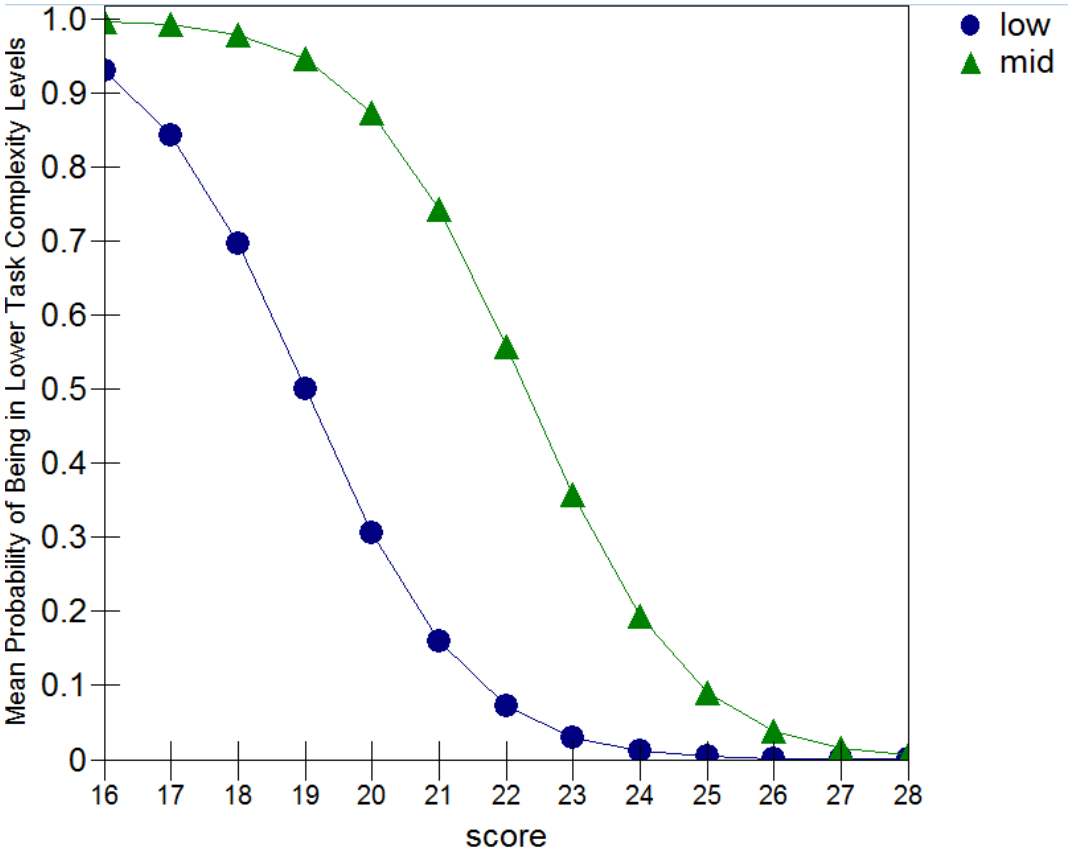


Figure 4.4. Mean probability of being in lower task complexity

Assuming that the between-interviewer variances in selecting task complexity at the cut-off score of each complexity level is not large, Figure *4.4* shows the general cumulative

probability of test takers' being given lower task complexity (*low* or *mid*) by interviewers. In the graph, the probability of choosing task prompts of a lower task complexity decreases as the score in the preceding task increases. This graph supports that the items in the OECT were adaptively selected by interviewers depending on test takers' performance score in the preceding task.

4.1.2. Variation of Rating Severity by Task Complexity

After verifying the adaptive context of the oral communication test, the use of the rating scale by experienced raters in the OECT was investigated based on a partial credit model (PCM) of the data with the many-facet Rasch measurement (MFRM). As the OPI section of the OECT included three tasks in a row and raters' evaluation of test takers' performance in the second or third round could be influenced by the scoring in the preceding task, the scores in the first round were initially analyzed before scores aggregated from all three rounds were analyzed. To generate the aggregated data, the scores from the three separate rounds were treated as independent scores from a single round.

Descriptive statistics for OECT scores

In the first round of rating, twenty-four raters rated 807 test takers using a holistic scoring rubric (see Appendix A) across three task complexity levels (*low*, *mid*, and *high*). Table 4.3 displays the descriptive statistics of ratings for each complexity level. The mean score of ratings for the three complexity levels ranged from 20.70 (*low*) to 22.62 (*high*). The standard deviation ranged from 1.58 (*high*) to 1.67 (*low*). The skewness values ranged from -0.96 (*high*) to 0.22 (*low*) and the kurtosis values ranged from -0.77 (*low*) to 0.63 (*high*).

Both the skewness and kurtosis values were within the acceptable range of ± 2.00 (Bachman, 2004), implying that the ratings of the three complexity levels were normally distributed. Table 4.4 displays the distribution of test taker scores by task complexity. The ratings ranged between 16 and 30, but the scores lower than 18 and higher than 25 were excluded, because they did not contain enough data points for the partial credit model and rating scale categories with less than 10 observations could have created some biased results when analyzed with a PCM (Eckes, 2015; Linacre, 2014). The *high* complexity level still had several score categories with less than 10 observations, such as score categories 18, 19, 21, and 25. Because there was not much data for the *high* complexity level, it is not possible to draw many conclusions from them.

Table 4.3. *Descriptive Statistics for Holistic Ratings in the First Round (N = 2,345)*

Task Complexity	N	Mean	SD	Min	Max	Skewness	Kurtosis
high	115	22.62	1.58	18	25	-0.96	0.63
mid	1,012	22.02	1.64	18	25	-0.37	-0.43
low	1,218	20.70	1.67	18	25	0.22	-0.77

Table 4.4. *Distribution of Test Taker Scores by Task Complexity in the First Round (N = 2,345)*

Task Complexity	Score							
	18	19	20	21	22	23	24	25
high	2	4	10	7	12	51	20	9
mid	23	54	124	151	209	275	126	50
low	112	214	273	213	195	161	40	10
Total	137	272	407	371	416	487	186	69

To overcome the potential bias that could be caused by the lack of observations per rating scale category, test taker scores from all three rounds were aggregated and used for

additional analysis. Table 4.5 presents the descriptive statistics for each complexity level with the aggregated data. The mean score of ratings for the three complexity levels ranged from 20.25 (*low*) to 23.13 (*high*) with standard deviations ranging from 1.67 (*low*) to 2.05 (*high*). The skewness values ranged from -0.71 (*high*) to 0.43 (*low*) and the kurtosis values ranged from -0.75 (*mid*) to 0.39 (*high*). Both the skewness and kurtosis values were within the acceptable range of ± 2.00 (Bachman, 2004). Table 4.6 displays the distribution of test taker scores for each task complexity level with the aggregated data, which contain enough observation points per rating scale category for a PCM (Eckes, 2015; Linacre, 2014).

Table 4.5. *Descriptive Statistics for Holistic Ratings with the Aggregated Data (N = 7,299)*

Task Complexity	N	Mean	SD	Min	Max	Skewness	Kurtosis
High	1,934	23.13	2.05	18	25	-0.71	0.39
Mid	3,144	21.42	1.84	18	25	0.02	-0.75
Low	2,221	20.25	1.67	18	25	0.43	-0.59

Table 4.6. *Distribution of Test Taker Scores by Task Complexity with the Aggregated Data (N = 7,299)*

Task Complexity	Score							
	18	19	20	21	22	23	24	25
high	21	48	89	164	299	663	301	349
mid	184	337	543	531	580	576	224	169
low	385	426	517	359	275	199	46	14
Total	590	811	1,149	1,054	1,154	1,438	571	532

Many-facet Rasch measurement

In order to examine how raters use the rating scales by different task complexity levels, multiple many-facet Rasch measurement (MFRM) models were analyzed. A task-complexity-related three-facet partial credit (PCM) MFRM model, which includes the facets

of test takers, raters, and task complexity, was used. While the rater and task complexity facets were centered at zero, the test taker facet was not centered to establish the origin of the measurement scale in logits. In addition, the test taker facet was set to be measured positively, which showed that high proficiency test takers had higher scores, and the rater and task complexity facets were set to be measured negatively, which showed that higher rater severity and higher task complexity measures indicated lower scores awarded to test takers. The main focus of the MFRM analyses was on the task difficulty and the rating scale use by comparing the task complexity-specific category threshold estimates (i.e., *Rasch-Andrich thresholds*) and standard error.

Assumptions of MFRM. There are multiple assumptions in the Rasch model: local independence of items, unidimensionality, absence of guessing, and equal item discrimination (Bond & Fox, 2015). First, the local independence of items (or ratings in this study) was measured by using Rasch-Cohen's Kappa and investigating the rating context. Second, the unidimensionality assumption of the construct in the test was investigated by investigating infit and outfit mean-square values. Finally, the assumptions of absence of guessing and equal item discrimination were assumed to be met in this study, because the rater training session was vigorous and raters generally were trained not to guess the scores in the test. Thus, only the local independence and unidimensionality assumptions were empirically tested.

Local independence. The local independence assumption, which requires that the quality of oral communication performance by test takers and ratings by experienced raters are independent of each other (Bond & Fox, 2015), was empirically tested by using the Rasch-Cohen's Kappa, which is (Observed agreement % - Expected agreement % / (100-

Expected agreement %) (Linacre, 2014). The Kappa of the data used in the first round was -0.067 [= (26.7%-31.3%)/(100-31.3%)], and the Kappa of the aggregated data was -0.061 [= (30.9%-34.9%)/(100-34.9%)]. These Kappa indices, which were close to 0, suggest that the independence assumption of the two datasets was sufficiently met. In addition, the vigorous rater training would indicate independence of rating in the first round although the aggregated scores of the all three rounds may have violated the independence assumption.

Unidimensionality. The unidimensionality assumption was tested by examining the infit and outfit statistics for task complexity. First, the infit mean-square values of *low*, *mid*, and *high* complexity levels in the first round of rating were 0.97, 1.01, and 1.01 and their outfit mean-square values were 0.97, 1.00, and 0.93. The infit mean-square values of *low*, *mid*, and *high* complexity levels in the aggregated data were 0.95, 1.00, and 1.03 and their outfit mean-square values were 0.94, 1.00, and 1.01. The observed mean-square values were almost identical to the Rasch model predicted values (= 1), implying that the unidimensionality assumption was met (Bond & Fox, 2015).

Ratings with the first-round data in OECT. The appropriate global fit statistics, a log-likelihood chi-square value, of the partial credit model for raters who considered task complexity in their scoring was 5,565.72 (df= 1,483, p=.00), indicating significant lack of global-data fit (Eckes, 2015; Linacre, 2012). While it is recommended that global fit be reported, this measure may be too sensitive to deviations from a best fitting model, particularly with large sample sizes (Eckes, 2015). The mean residual and the means standardized residual were both 0.00 and the standard deviation of the standardized residuals (S.D.) was 0.99, indicating that the estimation was successful (Linacre, 2014). This Rasch model predicted about 77.62% of the variance of the scores with 22.38% of residuals.

Table 4.7. *Rasch Measurement Summary Statistics with the First-Round Data*

	Test takers	Raters [(a)]	Task Complexity [(a)]
Measures			
Mean (SE)	-0.03 (0.77)	0.00 (0.17)	0.00 (0.06)
RMSE	0.81	0.19	0.07
Adjusted True SD	1.98	0.42	0.44
Infit Mean-Square			
Mean	0.92	0.94	1.00
Outfit Mean-Square			
Mean	0.92	0.94	0.97
Fixed chi-square (model)	5037.3**	558.4**	264.7**
df	806	23	2
Separation ratio (G)	2.44	2.26	6.24
Separation (strata) index (H)	3.58	3.34	8.65
Separation reliability (R)	.86	.84	.97

Note: [(a)]Rater and task complexity facets were centered at zero. $^{**}p < .01$.

Table 4.7 shows the summary Rasch statistics of the task-complexity-related three-facet partial credit model with the first-round data. The mean proficiency score of the test takers was -0.03 in logits, while the mean scores of the rater severity and the task complexity were 0, because they were centered. The infit mean-square values of the test taker, rater, task complexity facets were 0.92, 0.94, and 1.00. The outfit mean-square values of the three facets were 0.92, 0.94, and 0.97. The infit and outfit mean-squares were all within acceptable range (0.5-1.5) (Linacre, 2014), thus indicating the estimation with the proposed model can be deemed as successful. The fixed chi-square of the three facets (test takers: 5,037.3, raters: 558.4, and task complexity: 264.7) were all statistically significant, which means that at least two of test takers, raters, or task complexity levels did not share the same value of test taker ability, rater severity, and task complexity, respectively. The separation ratio G, which is the True SD divided by the root means-square error (RMSE), of the test taker facet was 2.44,

indicating the variability of the test taker ability was more than two times larger than the precision of the measure (Linacre, 2014). As scores lower than 18 and higher than 25 were excluded from the data, which thus had a heavy-tailed distribution, the separation (strata) index H of the test taker facet, instead of the separation ratio G, was used for the current analysis (Linacre, 2014). The strata H of the test taker facet was 3.58, indicating that the test could statistically distinguish between *high*, *mid*, and *low* proficiency levels of test takers (Linacre, 2014). The separation index H for rater and task complexity facets was 3.34 and 8.65, indicating that there were nearly three statistically distinct classes of rater severity and eight classes of task complexity, respectively. The reliability of the separation, which is the ratio of the true variance of the measures to the observed variance, for the test takers, raters, and task complexity were .86, .84, and .97, respectively.

Table 4.8 provides information about the rating scale difficulty of the three different complexity levels. The infit mean-squares of the *low, mid,* and *high* task complexity prompts were 0.97, 1.01, and 1.01. The outfit mean-square of the three task complexity levels were 0.97, 1.00, and 0.93. The infit and outfit mean-squares were all within an acceptable range (0.5-1.5) (Linacre, 2014). In the current model, the rating scale difficulties (and their standard errors) of the prompts with *low, mid,* and *high* task complexity levels were 0.61 (0.03), -0.16 (0.04), and -0.44 (0.11) on the logit scale, and they were statistically different from each other considering their standard errors. The high standard error for *high* task complexity may due to the lack of observed scores in each score category. *Low* task complexity was calculated as the most difficult one, contradicting the general expectation that *high* complexity tasks are generally the most difficult items and *low* complexity tasks are the easiest items. These unexpected outcomes suggest potential biases related to task complexity.

Table 4.8. *Rating Scale Difficulty and Infit and Outfit Mean-Squares with the First-Round Data*

Task Complexity	Difficulty (logits)		Infit Mean-Square	Outfit Mean-Square
	Measures	SE		
high	-0.44	0.11	1.01	0.93
mid	-0.16	0.04	1.01	1.00
low	0.61	0.03	0.97	0.97

Note: Raters and task complexity facets were centered at zero.

Table 4.9 shows the average measures and standard errors of the *Rasch-Andrich thresholds* (Bond & Fox, 2015), the transition point at which two adjacent categories are equally probable, of each category. The fit statistics of the three rating scales, ranging from 0.93 to 1.01, showed good fit within the acceptable range (0.5-1.5) (Linacre, 2014). The average measures of the test takers' ability, the last two columns in Table 4.9, increased with higher score categories and the difference between mean thresholds of rating scale categories was generally larger than the corresponding standard deviations, suggesting that the categories in the rating scale generally function appropriately (Eckes, 2015).

Table 4.9. *Thresholds of Each Task Complexity with the First-Round Data*

Score category	Low		Mid		High		Mean	
	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SD
19	-4.20	0.13	-3.66	0.26	-3.09	0.85	-3.65	0.56
20	-2.62	0.09	-2.81	0.15	-2.42	0.58	-2.62	0.20
21	-1.16	0.09	-1.38	0.12	-0.36	0.40	-0.97	0.54
22	-0.40	0.09	-0.48	0.11	-0.64	0.36	-0.51	0.12
23	0.64	0.11	0.79	0.10	-0.62	0.30	0.27	0.77
24	2.99	0.19	3.02	0.11	3.03	0.28	3.01	0.02
25	4.75	0.41	4.53	0.19	4.10	0.44	4.46	0.33

For a comparison of rating scale uses across different task complexity levels, the thresholds with the standard errors of the Rasch-Andrich thresholds of each category in Table

4.9 were compared. The standard error of the Rasch-Andrich thresholds of each category was used to calculate the statistical significance test between adjacent score categories, for example, when the threshold level from score category 21 to 22 with a *low* task complexity was -0.40 with the standard error of 0.09. Thus, the threshold level was -0.40 ± (1.96*0.09) with a 95% confidence interval. The comparison of thresholds from category 21 to 22 with *low* complexity tasks with *mid* complexity showed that the threshold with *low* complexity tasks [-0.40 ± (1.96*0.09)] was not statistically different from that with *mid* complexity tasks [-0.48 ± (1.96*0.11)], suggesting that a rating of 21 with *mid* complexity tasks may not be more difficult (or easier) for test takers to attain relative to a rating of 22. Only the rating scale with *high* complexity prompts demonstrated different rating thresholds than those with *mid* and *low* complexity prompts, which may be due to the lack of observed scores and/or uneven distribution of observed scores with each task complexity level. To address the potential biases caused by the lack of observed scores, another partial credit model with the aggregated data from all three rounds was analyzed.

Ratings with the aggregated data in OECT. Table 4.10 shows summary Rasch statistics of the task-complexity-related three-facet partial credit (PCM) MFRM model with the aggregated data from all three rounds. The chi-square value of the global data-model fit was 16,936.04 (df = 4,604, p = .00), indicating a significant lack of global-data fit (Eckes, 2015; Linacre, 2012). As the size of the sample is large, the lack of global-data fit does not necessarily mean that the model does not fit to the data (Eckes, 2015). As was evident in the MFRM model with the first-round data, the mean residual (=0.00), the mean standardized residual (=0.00), and the standard deviation of the standardized residuals (=0.99) show that the estimation with the proposed model can be claimed as successful (Linacre, 2014). This

Rasch model predicted about 79.25% of the variance of the scores. As shown in Table 4.10, the fit indices of the partial credit model with the aggregated data are almost equal to those with the first-round data. The mean proficiency score of the test takers was -0.05 on a logit scale, and the rater severity and task complexity were centered at 0.00. The infit mean-square values of the test taker, rater, and task complexity facets were 0.95, 0.94, and 0.95. The outfit mean-square values of the three facets were 0.95, 0.95, and 0.98. Both the infit and outfit mean-squares were all within the acceptable range (0.5-1.5) (Linacre, 2014).

Table 4.10. *Rasch Measurement Summary Statistics with the Aggregated Data*

	Test takers	Raters[(a)]	Task Complexity[(a)]
Measures			
Mean (SE)	-0.05 (0.77)	0.00 (0.10)	0.00 (0.02)
RMSE	0.82	0.11	0.03
Adjusted (True) SD	2.10	0.39	0.53
Infit Mean-Square			
Mean	0.95	0.94	0.95
Outfit Mean-Square			
Mean	0.95	0.95	0.98
Fixed chi-square (model)	15,779.1**	1255.0**	1170.6**
df	2418	23	2
Separation ratio (G)	2.56	3.49	21.30
Separation (strata) index (H)	3.74	4.98	28.73
Separation reliability (R)	.87	.92	1.00

Note: [(a)]Raters and task complexity facets were centered at zero. $^{**}p < .01$.

The fixed chi-square of 1170.6 with 2 degrees of freedom ($p = .00$) shows that at least two complexity levels were significantly different in terms of their difficulty. Table 4.11 shows that the rating scale difficulty levels (and standard errors) of the prompts with *low*, *mid*, and *high* task complexity levels were 0.61 (0.03), 0.09 (0.02), and -0.69 (0.03). The

reported separation (strata) index H was 28.73, with a reliability of 1.00, indicating that the three complexity levels could be statistically different considering the standard errors of the complexity measures. As was the model with the first-round data in Table 4.8, *low* task complexity was calculated as the most difficult while *high* complexity as the easiest, which does not meet the expectation that *high* complexity prompts are the most difficult tasks to test takers. The infit and outfit mean-square values of task complexity were between 0.94 and 1.03 and were within the acceptable range (0.5-1.5) (Linacre, 2014). The task complexity worked as the Rasch model expected with the aggregated data.

Table 4.11. *Rating Scale Difficulty and Infit and Outfit Mean-Squares with the Aggregated Data*

Task Complexity	Difficulty		Infit Mean-Square	Outfit Mean-Square
	Measures (logits)	SE		
high	-0.69	0.03	1.03	1.01
mid	0.09	0.02	1.00	1.00
low	0.61	0.03	0.95	0.94

Note: Raters and task complexity facets were centered at zero.

Table 4.12. *Outfit Mean-Square Values for Each Task Complexity with the Aggregated Data*

Score category	Low	Mid	High
18	1.1	1.0	0.8
19	1.0	1.0	1.1
20	0.8	0.9	1.0
21	0.9	1.0	1.0
22	0.9	0.9	1.0
23	0.9	1.0	0.9
24	1.2	1.2	1.0
25	1.2	1.1	1.1

Average measures and the *outfit mean-square* values of each rating scale were investigated to see if the rating scale functioned appropriately. As shown in Table 4.12, the

outfit mean-square values ranged from 0.8 to 1.2, which were within the acceptable range of 0.5 to 1.5 (Linacre, 2014). In addition, Table 4.13 shows the average measures and standard errors of the *Rasch-Andrich thresholds* of each category with the aggregated data. Unlike the thresholds with the first-round data in Table 4.9, the average measures of the test takers' ability with the aggregated data seem to generally increase with higher score categories across the three rating scales, and the thresholds were generally consistent across three rating scales of three task complexity levels.

Table 4.13. *Thresholds of Each Task Complexity with the Aggregated Data*

Score	Low		Mid		High		Mean	
category	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SD
19	-3.73	0.08	-3.70	0.10	-3.02	0.27	-3.48	0.40
20	-2.46	0.06	-2.49	0.07	-1.96	0.16	-2.30	0.30
21	-1.04	0.07	-1.12	0.06	-1.30	0.12	-1.15	0.13
22	-0.30	0.08	-0.27	0.06	-0.56	0.09	-0.38	0.16
23	0.67	0.09	0.90	0.06	0.21	0.07	0.59	0.35
24	2.88	0.17	2.91	0.08	2.92	0.07	2.90	0.02
25	3.97	0.33	3.76	0.12	3.71	0.10	3.81	0.14

Based on the findings with partial credit analyses with both the first-round and the aggregated data, as shown in Table 4.8 and Table 4.11, test takers who were administered more complex tasks tend to receive higher scores, which is contradictory to what is generally expected. Even after addressing the potential bias in the partial credit model with the first-round data due to the lack of observed scores, the unexpected rating scale difficulty has not changed. This discrepancy could be due to (a) test takers' performance difference with different complexity level tasks and/or (b) raters having adjusted their ratings depending on the task complexity levels.

4.1.3. Section Summary

This section presented the results of statistical analyses that examined the effect of test takers' performance on the interviewer's selection of the complexity level of the following tasks and the effect of task complexity on rater severity in a performance-based L2 oral communication test. First, the interviewer's selection of task complexity levels, which tested the adaptiveness of the OECT, was examined with the multilevel ordinal regression analysis. The random slope cumulative logit model (Model 3) explained that the effect of test takers' performance, or score, on the selection of task complexity levels was generally consistent across interviewers. The random intercept cumulative logit model (Model 2), which relaxed the random slope assumption, further showed the effect of the mean probability of selecting tasks in lower task complexity. Based on the results of the two multilevel ordinal regression models, the adaptiveness of the OECT was confirmed.

In order to examine the effect of task complexity on rater severity in the operational data of OECT, this section reported and discussed the results of partial credit models of MFRM analyses. The thresholds of each task complexity were initially checked and it was found that they were generally consistent across the rating scales of three task complexity levels. The average task difficulty of the three levels of task complexity was then analyzed. The *low* complexity prompts were statistically calculated as the most difficult item while the *high* complexity prompts as the easiest ones. This unexpected finding presents the possibility of either test takers' performance difference or rater severity change depending on task complexity.

4.2. Effect of Task Complexity on Test Takers' Linguistic Outputs and Proficiency Scores (Sub-Study 1)

To further understand the threshold discrepancy described in Section 4.1.2, test takers' performance differences with *high* and *low* prompt complexity were examined by comparing their linguistic outputs and proficiency scores. Eighty-one test takers who were given prompts on all three complexity levels were selected from the OECT dataset for linguistic output and 40 of test takers were chosen for proficiency scores analyses. This section answers to the questions RQ 1-2 and RQ 1-3, on whether or not the threshold discrepancy was due to the variation of test takers' linguistic outputs.

4.2.1. Effect of Task Complexity on Test Takers' Linguistic Outputs

Descriptive statistics of linguistic complexity and fluency

Table 4.14 and

Table 4.15 display the descriptive statistics of the measures of linguistic complexity and fluency by task complexity. The linguistic complexity measures correspond to lexico-grammar scores in the analytic scoring by human judges; the fluency measures correspond to the fluency scores. These linguistic outputs were measured in terms of their frequency.

Table 4.14. *Descriptive Statistics of the Linguistic Complexity Measures (N = 81)*

Measures	Complexity Level	Min.	Max.	Median	Interquartile Range
Subordinate Index	High	0.00	4.50	0.90	0.91
	Low	0.00	6.00	0.80	0.83
Guiraud Advanced 1000	High	0.00	1.29	0.52	0.42
	Low	0.00	1.53	0.45	0.50

Wilcoxon signed-rank test results of linguistic outputs

In order to examine the linguistic output differences by test takers who performed on *high* and *low* complexity prompts, the Wilcoxon's signed-rank test, which is a non-parametric alternative to a paired samples *t*-test, was conducted. This approach was deemed appropriate because the frequency of linguistic outputs was not normally distributed (Field, 2009).

Table 4.15. *Descriptive Statistics of Fluency Measures (N = 81).*

Measures	Complexity Level	Min.	Max.	Median	Interquartile Range
Pruned Speech Rate (Pruned syllables / seconds)	High	0.44	4.76	2.73	1.07
	Low	0.99	4.81	2.92	1.06
Unpruned Speech Rate (No. of syllables / seconds)	High	0.69	4.95	3.09	1.07
	Low	1.38	4.85	3.25	0.93
Mean length of runs (No. of syllables / runs)	High	2.79	160	10.08	11.70
	Low	3.22	145	12.40	14.53
Phonation time ratio (Phonation time / Total Length)	High	0.61	1.00	0.84	0.14
	Low	0.53	1.00	0.85	0.13
Repairs per AS-unit	High	0.00	3.00	0.63	0.67
	Low	0.00	5.00	0.50	0.55
Filled Pauses (uh, uhm) per AS-unit	High	0.00	4.67	0.75	1.10
	Low	0.00	11.00	0.80	1.39
Preparation Time (seconds)	High	0.76	14.88	3.55	3.72
	Low	0.09	9.87	2.18	2.03

Assumptions of Wilcoxon signed-rank test. The Wilcoxon signed-rank test needs to meet the following assumptions: a) minimum number of cases should be five pairs, b) each pair should be random and independent, and c) the distribution of the frequency differences in each pair should be symmetric (Rey & Neuhäuser, 2011). As there were 81 pairs of

observations in the data, the minimum case assumption was met. The 81 test takers took the test individually, thus making the frequency scores with those test takers' audio files random and independent. Finally, the boxplots in Appendix E display that the frequency differences of the nine linguistic indices in Table 4.16 and Table 4.17 follow a symmetrical distribution. Thus, it can be argued that the dataset used to examine the lexico-grammar and fluency features met the assumptions of the Wilcoxson signed-rank test.

Linguistic output analysis. The results of the Wilcoxon signed-rank tests in Table 4.16 showed that the *Subordinate Index* for *low* complexity (*Mdn* = 0.80) was not significantly different from that for *high* complexity (*Mdn* = 0.90, z = -0.84, p = .40, r = -.07). In addition, *Guiraud Advanced 1000* for *low* complexity (*Mdn* = 0.45) was not significantly different from *high* complexity (*Mdn* = 0.52, z = -1.21, p = .23, r = -.10). Both the *Subordinate Index* and *Guiraud Advanced 1000* failed to reject the null hypothesis that the linguistic outputs with *low* complexity and those with *high* complexity are the same.

Table 4.16. *Wilcoxon Signed-Rank Test of Linguistic Complexity Measures (Low–High Complexity)*

Measures	$Z^{(a)}$	$p^{(b)}$	$r^{(c)}$
Subordinate Index	-0.84 $^{+}$	.40	-0.07
Guiraud Advanced 1000	-1.21 $^{+}$	.23	-0.10

Note. [(a)]Wilcoxon signed-rank test; [b]Asympt. sig (2-tailed): $^{**}p < .01$ $^{*}p < .05$; [(c)]Effect size. r = .10 is a small effect; r = .30 is a medium effect; and r = .50 is a large effect (Cohen, 1992, p. 157); z^{+} is based on positive ranks, z^{-} is based on negative ranks.

The Wilcoxon signed-rank tests with fluency measures, in contrast, showed that there were significant differences in *Phonation-Time Ratio*, *Preparation Time,* and *Repairs per AS-*

Unit among the seven fluency measures displayed in Table 4.17. The *Phonation-Time Ratio* for *low* complexity (*Mdn* = 0.85) was significantly higher than that for *high* complexity (*Mdn* = 0.84, z = -2.65, p = .01, r = -.21). In addition, the *Preparation Time* for *low* complexity (*Mdn* = 2.18) was significantly shorter than that for *high* complexity (*Mdn* = 3.55, z = -4.79, p = .00, r = -.38). The results of the *Phonation-Time Ratio* and *Preparation Time* analyses suggest that more complex prompts tend to prompt test takers' relatively shorter utterance time, but promote longer preparation time. The result of the *Repairs per AS-Unit* test indicated that the frequency of repairs with *low* complexity prompts (*Mdn* = 0.50) was significantly higher than that with *high* complexity prompts (*Mdn* = 0.63, z = -2.06, p = .04, r = -.16), which indicates that test takers repaired their speech more frequently with more complex prompts. Even though the effect size is relatively small (r = -.16), the frequency of repairs, coupled with preparation time, might have influenced the phonation time.

Table 4.17. *Wilcoxon Signed-Rank Test of Fluency Measures (Low–High Complexity)*

Measures	$z^{(a)}$	$p^{(b)}$	$r^{(c)}$
Pruned Speech Rate	-0.92 $^{+}$	.36	-0.07
Unpruned Speech Rate	-1.00 $^{+}$	.32	-0.08
Mean length of runs (script)	-1.72 $^{+}$	.09	-0.13
Phonation time ratio **	-2.65 $^{+}$	.01**	-0.21
Repairs per AS-unit*	-2.06 $^{-}$	.04*	-0.16
Filled Pauses (uh, uhm) per AS-unit	-0.74 $^{+}$	.46	-0.06
Preparation Time (seconds) **	-4.79 $^{-}$	.00**	-0.38

Note. [(a)]Wilcoxon signed-rank test; [(b)]Asympt. sig (2-tailed): $^{**}p < .01$ $^{*}p < .05$; [(c)]Effect size. r = .10 is a small effect; r = .30 is a medium effect; and r = .50 is a large effect (Cohen, 1992, p. 157); z^{+} is based on positive ranks, z^{-} is based on negative ranks.

4.2.2. Effect of Task Complexity on Test Takers' Proficiency Scores

Descriptive statistics of proficiency scores

Table 4.18 shows the descriptive statistics of the mean and standard deviation of the lexico-grammar and proficiency scores graded by human raters with the converted analytic (lexico-grammar and fluency) scoring rubric (scores of 0-13; see Appendix A). The first two columns provide the scoring category and the complexity level of the interview prompts. The mean (and its standard deviation) of lexico-grammar scoring for *high* task complexity was 7.36 (1.78) and that of *low* task complexity was 7.71 (1.94). The mean (and its standard deviation) of fluency scoring was 6.96 (2.16) and that of *low* task complexity was 7.06 (2.09).

Table 4.18. *Descriptive Statistics of the Scores by Human Judges (N = 40)*

Scoring Category	Complexity Level	Mean	Standard Deviation
Lexico-grammar	High	7.36	1.78
	Low	7.71	1.94
Fluency	High	6.96	2.16
	Low	7.06	2.09

Note: Analytic scales range between 0 and 13.

Paired samples *t*-test for proficiency scores

In order to examine the lexico-grammar and fluency score differences when test takers performed with *high* and *low* task complexity prompts, two paired samples *t*-tests were conducted. The main focus of this paired samples analysis was on the raters' analytic scoring differences between the *high* and the *low* complexity prompts.

Assumptions of paired samples *t*-test. The paired samples *t*-test also follows the general parametric test assumptions: a) the dependent variables should be continuous data, b) there should be no significant outliers, and c) the sampling distribution of the score differences should be normal. First, as the analytic scoring was conducted on a 0-13 scale, the score differences between *high* and *low* task complexity were assumed to be continuous data. Second, no significant outliers were found in the box plot as shown in Figure *4.5*. Finally, the normality of the sampling distribution of the score differences was examined using the Kolmogorov-Smirnov (K-S) test. The lexico-grammar scores, $D(40) = 0.11$, $p = .20$, and the fluency scores, $D(40) = 0.11$ $p=.20$, both failed to reject the null hypothesis that the sampling distribution was normal. Unlike the independent samples *t*-test, the assumptions of independence and homogeneity of variance do not need to be tested, because scores for the paired samples *t*-test are derived from the same groups of people (Field, 2009).

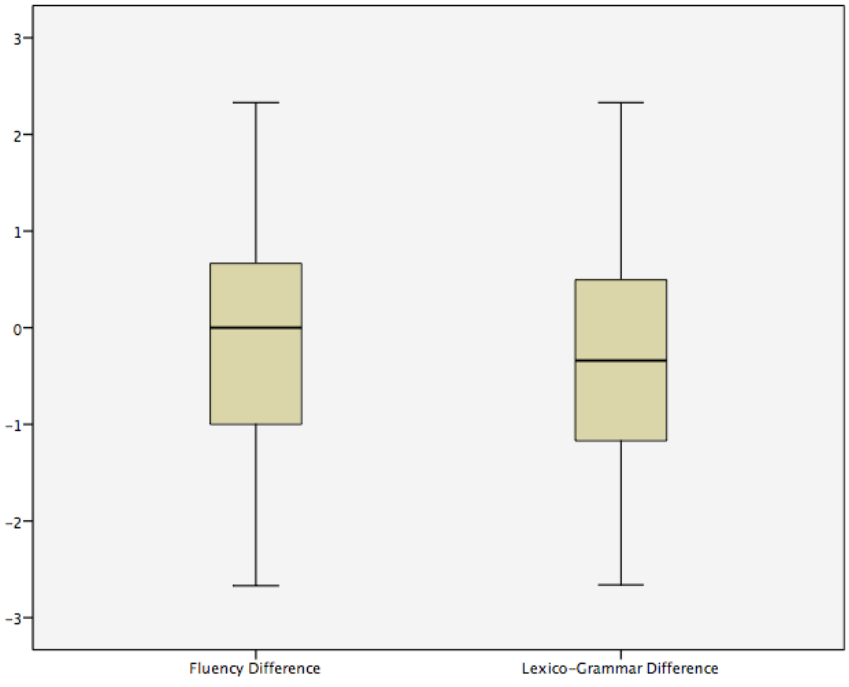


Figure 4.5. Boxplot of analytic scoring

Analytic score differences. As shown in Table 4.19, the results of the paired samples *t*-tests indicate that there were no statistically significant proficiency score differences for *high* and *low* prompt complexities, lexico-grammar score: $t(39) = 1.92$, $p = .06$, fluency: $t(39) = 0.50$, $p = .62$, which suggests that those two scores were not likely to be influenced by prompt complexity.

Table 4.19. *Paired Samples t-test (Within-Subjects Design) with High and Low Levels (N=40)*

Scoring Category	Paired Differences					
	Mean	*SD*	*SE*	*t*	*df*	*p*
Lexico-grammar (Low – High)	0.35	1.16	0.18	1.92	39	.06
Fluency (Low – High)	0.10	1.25	0.18	0.50	39	.62

Note: Analytic scales range between 0-13.

While the Wilcoxon signed-rank tests in the previous section showed that there were significant differences in the fluency measures, such as *Phonation-Time Ratio*, *Preparation Time*, and *Repairs per AS-Unit*, these differences were not translated into the fluency score differences measured by human raters. This discrepancy between linguistic measures and human scoring may suggest that human raters consider task complexity when evaluating test takers' performance; this is because there could be human scoring differences that were affected by fluency measure differences if the size of fluency measure differences was big enough to be recognized, but was not selectively disregarded by human raters because of the task complexity difference.

4.2.3. Section Summary

This section presented the results of statistical analyses that examined the effect of task complexity on test takers' linguistic outputs and proficiency scores. Multiple Wilcoxon's signed-rank tests were conducted and revealed that the linguistic complexity and fluency measures, except *Phonation-Time Ratio*, *Repairs per AS-Unit* and *Preparation Time,* did not show any statistical differences by task complexity. The statistical differences in the fluency measures, which are related to preparation time, suggest that test takers spent more time when they were given difficult prompts. Multiple paired samples *t*-tests were also conducted to find the score difference in the lexico-grammar and fluency scores by task complexity, but there were no statistically different results. This discrepancy between the difference of linguistic measures and that of human scoring may suggest that human raters considered task complexity during the scoring in an adaptive performance-based oral communication context.

4.3. Quantitative Analysis of Rating Scales Use (Sub-Study 2)

This section reports on the findings of the second research question RQ 2: How do raters adjust their score assignment depending on their understanding of prompt complexity in a performance-based oral communication test? The results in the preceding sections indicated how raters used their rating scales in an adaptive performance-based oral communication assessment setting. This section focuses more on raters' rating scale use by controlling the adaptiveness of the test. Based on the raters' preference toward the consideration of task complexity in their scoring of the oral communication data, separate MFRM analyses were conducted on raters who considered or did not consider the complexity

in oral communication rating. Raters who answered, in their interview with the interview question 2-c (see Appendix D), that they took the task complexity into account when scoring test takers' performance were categorized as the raters in the *scale adjusting group* [hereafter referred to as "Scale Adjusting (SA)" group], while those who answered that they did not consider task complexity in their scoring as the raters in the *scale non-adjusting group* [hereafter referred to as "Scale Non-Adjusting (SNA)" group].

4.3.1. Quantitative Analysis of Rating Scale Use

After controlling the adaptiveness of the prompt selection, raters' rating scale use with or without the knowledge of the task complexity was investigated based on a partial credit model (PCM) of the data with the many-facet Rasch measurement (MFRM).

Descriptive statistics for 40 test taker data

Nine newly trained raters judged 80 audio files of 40 test takers, who produced speech from *high* and *low* complexity prompts, using the converted holistic scoring rubric (scores of 0-13; see Appendix A). These raters were different than the ones that rated the operational test. Table 4.20 displays the descriptive statistics of ratings for each complexity level with or without raters' understanding of task complexity. When raters rated without knowing the complexity levels of the prompt (hereafter referred to as "Answer Only" rating context), the mean ratings ranged from 6.43 (*high* complexity) to 7.05 (*low* complexity), and the standard deviation ranged from 2.74 (*high* complexity) to 2.62 (*low* complexity). When raters rated with the knowledge of the complexity levels of the prompt (hereafter referred to as "Question and Answer" rating context), the mean ratings ranged from 6.72 (*high*

complexity) to 6.76 (*low* complexity) and the standard deviation ranged from 2.51 (*high* complexity) to 2.40 (*low* complexity). Skewness and kurtosis values ranged from -0.69 (kurtosis; low complexity; without prompt knowledge) to 0.21 (skewness; high complexity; with prompt knowledge), and were all within the acceptable range of ± 2.00 (Bachman, 2004), implying that the ratings were normally distributed. As there was a small number of score categories 0, 1, and 13 to run a partial credit model, scores of 0, 1, and 2 were combined and scores of 12 and 13 were combined.

Table 4.20. *Descriptive Statistics for the New Raters' Holistic Ratings*

Rating Context	Task Complexity	N	Mean	SD	Min	Max	Skewness	Kurtosis
Answer Only	High	360	6.43	2.74	0	13	0.06	-0.56
	Low	360	7.05	2.62	1	13	-0.02	-0.69
Question and Answer	High	360	6.72	2.51	1	13	0.21	-0.51
	Low	360	6.76	2.40	1	13	0.16	-0.17

Many-facet Rasch measurement

In order to examine how raters used the rating scales depending on their knowledge of the task complexity, two separate analyses of the partial credit many-facet Rasch measurement (MFRM) model were conducted. The first MFRM model was analyzed for the raters in the SA group. The second model was analyzed for the raters in the SNA group. The main focus of these MFRM analyses was on the rating scale use by two groups of raters, who indicated they did or did not consider task complexity in their scoring. The task complexity-specific category threshold estimates and their standard errors (i.e., *Rasch-Andrich thresholds*) and task difficulty of the prompts with different task complexity levels were

compared between the two groups of raters and among *high*, *mid*, and *low* task complexity prompts.

Assumptions of MFRM. As was discussed in Section 4.1.2, among the assumptions of the Rasch model (i.e., local independence of items, unidimensionality, absence of guessing, and equal item discrimination [Bond & Fox, 2015]), the local independence and unidimensionality assumptions were empirically tested.

Local independence. The local independence assumption was empirically tested using the Rasch-Cohen's Kappa, which is (Observed agreement % - Expected agreement % / (100 - Expected agreement %) (Linacre, 2014). The Kappa of the data used in the SA group raters was -0.044 [= (14.8%-18.4%)/(100-18.4%)], and the Kappa of the data in the SNA group raters was -0.019 [= (19.0%-20.5%)/(100-20.5%)], implying that the independence assumption of the two datasets was sufficiently met. Vigorous rater training, randomization of the rating order, and enough of a break, which is more than five days, between the first rating without the prompt knowledge and the second rating with the prompt knowledge, would indicate independence of rating scores.

Unidimensionality. The unidimensionality assumption was tested by examining the infit and outfit statistics for task complexity. First, the infit mean-square values of *low*, *mid*, and *high* complexity levels with the SA group raters were 0.99, 0.95, and 1.09 and their outfit mean-square values were 1.04, 0.94, and 1.06, respectively. The infit mean-square values of *low*, *mid*, and *high* complexity levels with the SNA group raters were 0.90, 0.96, and 1.06 and their outfit mean-square values were 0.93, 1.04, and 1.06, respectively. The observed mean-square values were almost identical to the Rasch model predicted values (=1.00), implying that the unidimensionality assumption was met (Bond & Fox, 2015).

Ratings by Scale Adjusting Group raters

Main effects. The appropriate global fit statistics, a log-likelihood chi-square value, of the MFRM model for raters in the SA group, or scale adjusting raters, was 3034.23 (*df* = 861, *p* = .00), indicating a significant lack of global data fit (Eckes, 2015; Linacre, 2012). This lack of global data fit does not necessarily mean that the model did not fit the data because global fit statistics are sensitive to sample sizes (Eckes, 2015). As the mean residual and the mean standardized residual were 0.00 and the standard deviation of the standardized residuals was 1.00, the estimation with the proposed model can be considered successful (Linacre, 2014). This Rasch model predicted about 76.92% of the variance of the scores with 23.08% of residuals.

Table 4.21. *Rasch Measurement Summary Statistics (SA Group)*

	Test takers	Raters[(a)]	Task Complexity[(a)]	Rating Context[(a)]
Measures				
Mean (SE)	-0.13 (0.16)	0.00 (0.06)	0.00 (0.05)	0.00 (0.04)
RMSE	0.16	0.06	0.05	0.04
Adjusted (True) SD	1.28	0.56	0.11	0.00
Infit MS				
Mean	1.01	1.01	1.00	1.01
Outfit MS				
Mean	1.01	1.01	1.00	1.01
Fixed chi-square (model)	2310.5**	476.4**	18.1**	1.2
df	39	5	2	1
Separation ratio (G)	7.88	8.96	2.31	0.00
Separation (strata) index (H)	10.84	12.28	2.42	0.33
Separation reliability (R)	.98	.99	.84	.00

Note. [(a)]Raters, task complexity, and rating context variables were centered at zero. **$p < .01$.

Table 4.21 shows the summary Rasch statistics of the three-facet partial credit (PCM) MFRM with the raters in the SA group. The mean proficiency score of the test takers was -0.13 in logits, but the rater severity, task complexity, and rating context were centered at 0.00. The infit mean-squares of the test taker, rater, task complexity, and rating context facets were 1.01, 1.01, 1.00, and 1.01 and their corresponding outfit mean-squares were 1.01, 1.01, 1.00, and 1.01. The infit and outfit mean-squares were all within the acceptable range (0.5-1.5) (Linacre, 2014), indicating the estimation with the proposed model was successful.

The fixed chi-square values of the test takers, raters, and task complexity were statistically significant, which means that at least two levels of test takers, raters, or task complexity level did not share the same value of test taker ability, rater severity, and task complexity, respectively. The separation ratio G and the strata index H of the test takers were 7.88 and 10.84, indicating that the test could statistically distinguish between more than eight proficiency levels of test takers (Linacre, 2014). The strata index H for rater and task complexity facets were 12.28 and 2.42, indicating that there were nearly 12 statistically distinct classes of rater severity and about three classes of task complexity, respectively. The reliability of the separation for the test takers, raters, and task complexity level were .86, .84, and .97.

Figure *4.6* presents the Wright map resulting from the partial credit analysis of the 40 test takers with two task complexity levels by six newly trained raters in the SA group. The first column in the Wright map provides a logit scale and the second column, test taker, shows the ability of test takers. The third column presents rater severity of the six raters in the SA group. As shown in Table 4.22, the infit and outfit mean-square values of the raters

were between 0.82 and 1.19, and they were all within the acceptable range (0.5-1.5) (Linacre, 2014), indicating that raters behaved as the Rasch model expected.

Table 4.22. *Rater Severity and Infit and Outfit Mean-Squares (SA Group)*

Raters	Rater Severity		Infit	Outfit
	Measures (logits)	SE	Mean-Square	Mean-Square
R2	0.74	0.06	0.82	0.81
R6	0.70	0.06	1.11	1.06
R5	-0.03	0.06	0.89	0.92
R4	-0.15	0.06	1.14	1.12
R3	-0.50	0.06	0.97	0.95
R9	-0.76	0.06	1.14	1.19

The fourth column provides information about the rating context how raters judged the test takers' audio files. The rating scale difficulty levels (and standard errors) of the prompts in the *Question and Answer* rating context and in the *Answer Only* rating context were 0.03 (0.04) and -0.03 (0.04) in logits, indicating that the average rating difficulty in both rating contexts was not statistically different.

Table 4.23. *Task Difficulty and Infit and Outfit Mean-Squares (SA Group)*

Task Complexity	Difficulty		Infit	Outfit
	Measures (logits)	SE	Mean-Square	Mean-Square
high	0.01	0.04	1.06	1.03
mid	0.14	0.05	0.94	0.93
low	-0.15	0.04	0.99	1.03

The fifth column provides information about the rating scale difficulty of the three task complexity levels (*low, equal,* and *high*). As shown in Table 4.23, the rating scale difficulty levels (and standard errors) of the prompts with *low, equal* and *high* task complexity were -0.15 (0.04), 0.14 (0.05) and 0.01 (0.04) in logits and their infit and outfit mean-square values

ranged from 0.93 (outfit mean-square; *mid* task complexity) to 1.06 (infit mean-square; *high* task complexity), indicating that the three task complexity levels met the expectations of the measurement model. As shown in Table 4.21, the reported separation ratio G and the strata H were only 2.31 and 2.42, with a reliability of .84, indicating that at least one of the task complexity categories was statistically different from one other complexity categories in terms of their difficulty. The last six columns provide information about the score categories for the rating scales of the three task complexities in the two rating contexts.

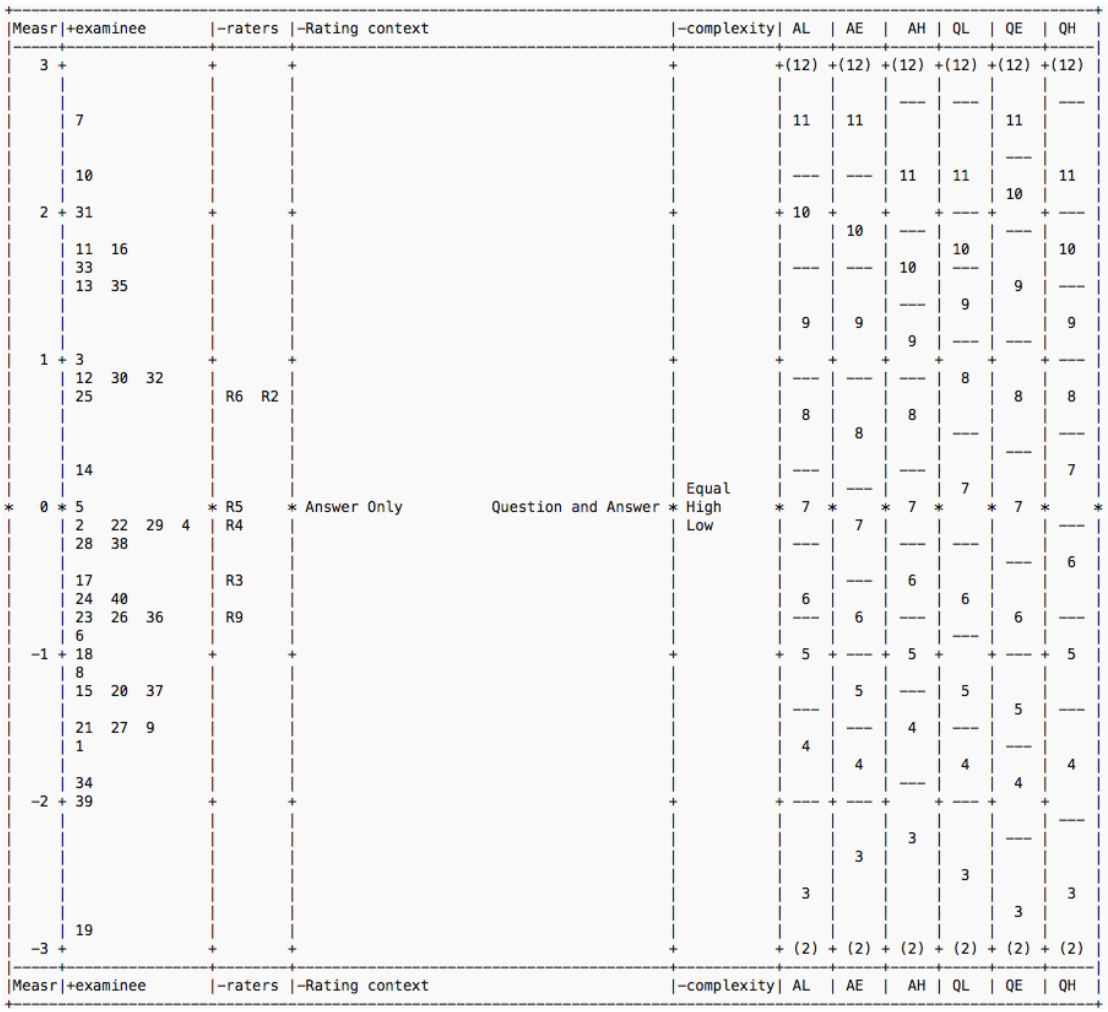


Figure 4.6. Wright map from raters who considered task complexity (SA group).

Note. AL = low task complexity in an answer only context, AE = equal task complexity in an answer only context, AH = high task complexity in an answer only context, QL = low task complexity in a question and answer context, QE = equal task complexity in a question and answer context, QH = high task complexity in a question and answer context.

Thresholds. Table 4.24 shows the average measures and standard errors (SE) of the *Rasch-Andrich thresholds* of the six ratings [three (*low*/*equal*/*high*) task complexity levels * two rating contexts (with/without knowing the prompts)]. For the comparison of rating scale uses across different task complexity levels in different rating context, *Rasch-Andrich thresholds* of these rating scales were compared. The standard error of the Rasch-Andrich thresholds of each category was used to calculate the statistical significance test between adjacent score categories with ± 1.96 * SE with a 95% confidence interval.

Table 4.24. *Rasch-Andrich Thresholds of the Rating Scale for Raters (SA Group)*

	AL		AE		AH		QL		QE		QH	
Score	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SE
3	-3.42	0.58	-2.38	0.39	-2.25	0.38	-2.58	0.49	-2.72	0.44	-2.60	0.47
4	-1.38	0.29	-1.32	0.36	-1.71	0.28	-1.09	0.32	-1.78	0.37	-2.16	0.30
5	-1.03	0.25	-1.89	0.35	-1.10	0.24	-2.27	0.28	-1.97	0.34	-0.94	0.24
6	-1.17	0.23	-0.94	0.32	-0.94	0.23	-0.62	0.23	-0.84	0.31	-1.01	0.22
7	-0.23	0.23	-0.45	0.31	-0.06	0.24	-0.43	0.22	-0.62	0.30	0.06	0.23
8	0.30	0.24	-0.14	0.32	0.26	0.25	0.15	0.23	0.47	0.33	0.78	0.25
9	0.56	0.25	1.00	0.37	0.86	0.27	2.34	0.28	0.80	0.37	0.87	0.28
10	2.07	0.29	1.99	0.41	1.30	0.30	0.74	0.31	2.51	0.42	1.13	0.30
11	2.29	0.35	1.83	0.44	1.92	0.33	2.41	0.34	2.87	0.49	2.90	0.35
12	2.01	0.42	2.30	0.55	1.71	0.38	1.34	0.36	1.29	0.54	0.97	0.38

Note. AL = low task complexity in an answer only context, AE = equal task complexity in an answer only context, AH = high task complexity in an answer only context, QL = low task complexity in a question and answer context, QE = equal task complexity in a question and answer context, QH = high task complexity in a question and answer context.

As shown in Figure *4.6* and Table 4.24, the *Rasch-Andrich thresholds* were only slightly different across six rating scales. For example, the Rasch-Andrich threshold from score 8 to 9 was 0.56 (95% CI: 0.07, 1.05) when raters judged test takers' performance with *low* complexity prompts without knowing the prompts (*AL*), while the same threshold with *high* complexity prompts (AH) was 0.86 (95% CI: 0.33, 1.39), indicating that they are statistically different from each other with a 95 % confidence interval. However, Figure *4.7* and Figure *4.8* show that confidence intervals with *high* complexity prompts in the two rating

contexts (with/without knowing the prompt), and there were many overlapping confidence intervals with *high* (*AH* vs. *QH*) and *low* (*AL* vs. *QL*) complexity prompts. As most confidence intervals overlapped with each other, it cannot be claimed that a threshold in one scale is different from a corresponding threshold in another scale.

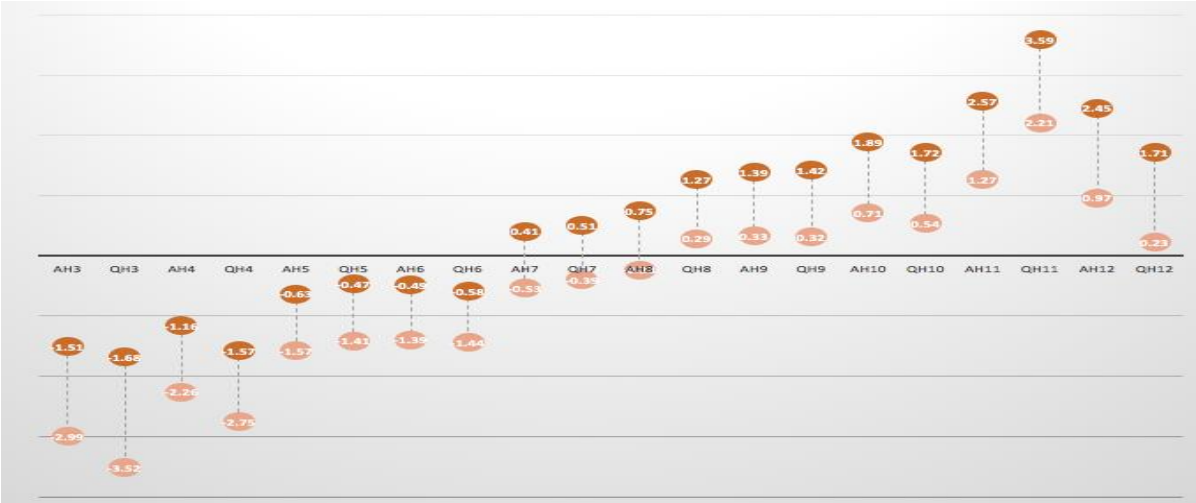


Figure 4.7. 95% Confidence intervals with *AH* and *QH* ratings.

Note. AH = high task complexity in an answer only context, QH = high task complexity in a question and answer context.

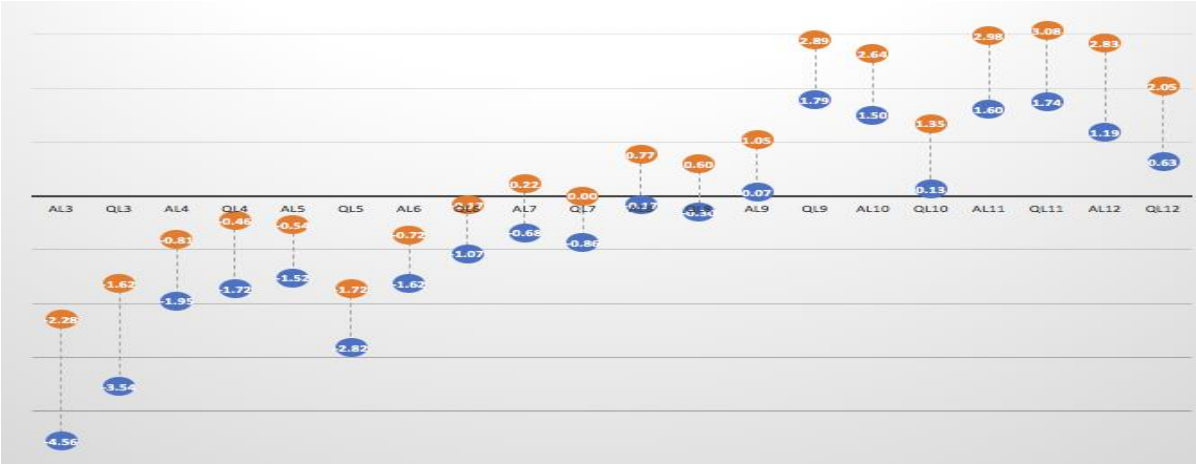


Figure 4.8. 95% Confidence intervals with *AL* and *QL* ratings.

Note. AL = low task complexity in an answer only context, QL = low task complexity in a question and answer context.

There were some thresholds that were different from their corresponding thresholds in another scale; for example, the threshold from 8 to 9 with *low* complexity prompts in the *Question and Answer* context (QL) was 2.34, which is statistically different from those in AL (0.56) and QH (0.87) considering their standard errors. This difference might be due to the reversed thresholds from 8 to 9 created in the QL rating scale. Reversed thresholds are sometimes created due to the lack of observed scores in each score category. The occurrence of reversed thresholds, however, does not necessarily mean that the order of the response categories is violated (Wetzel & Carstensen, 2014). Thus, instead of comparing individual thresholds with confidence intervals, interactions between task complexity and rating contexts were analyzed to provide a more general picture of the effect of task complexity on the change of rater severity.

Task complexity interaction with rating context. To estimate the size of any rating context (i.e., *Answer Only* and *Question and Answer* rating context) effect on the rating scale difficulty, the interaction of task complexity (i.e., *high*, *equal*, and *low*) with rating contexts was analyzed. The interaction model included rating context and task complexity as dummy facets to investigate the interactions between rating context and task complexity. Table 4.25 shows the pairwise bias report for raters in the SA group who answered that they considered task complexity in their rating. *High* complexity tasks were 0.31 logits more difficult in the *Answer Only* rating context than those in the *Question and Answer* rating context with a 99% confidence level (t[365] = 3.70, p = .00). In contrast, *low* complexity tasks were 0.14 logits easier in the *Answer Only* context than those in the *Question and Answer* rating context with a 90% confidence level (t[365] = -1.71, p = .09). *Equal* complexity tasks did not show any

statistical differences in task difficulty depending on the *Answer Only* and *Question and Answer* rating contexts.

Table 4.25. *Pairwise Bias Report for Task Complexity with Rating Context (SA Group)*

	Answer Only		*Question and Answer*				*Welch's t*			
Task Complexity	*Measure*	*SE*	*Measure*	*SE*	*Contrast*	*SE*	*t*	*df*	*p*	$d^{(a)}$
high	0.20	0.06	-0.11	0.06	0.31	0.08	3.70	365	.00**	0.19
equal	0.09	0.08	0.14	0.08	-0.04	0.11	-0.39	221	.69	-0.03
low	-0.19	0.06	-0.04	0.06	-0.14	0.08	-1.71	365	.09	-0.09

Note. [(a)]Effect size (*d*) was divided by two because Welch's t-test in the current study was a paired test (Dunlap et al., 1996). Effect size *d* = .20 is a small effect; *d* = .50 is a medium effect; and *d* = .80 is a large effect (Cohen, 1992, p. 157). ***p* < .01, **p* < .05.

Table 4.26. *Pairwise Bias Report for Rating Context with Task Complexity (SA Group)*

	Task 1			*Task 2*					*Welch's t*			
Rating Context	*Measure*	*SE*	*Complexity*	*Measure*	*SE*	*Complexity*	*Contrast*	*SE*	*t*	*df*	*p*	$d^{(a)}$
Answer Only	-0.19	0.06	low	0.09	0.08	equal	-0.28	0.10	-2.90	254	.00**	-0.18
	-0.09	0.06	low	0.20	0.06	high	-0.38	0.08	-4.69	365	.00**	-0.25
	0.19	0.08	equal	0.20	0.06	high	-0.10	0.10	-1.07	251	.29	-0.07
Question and Answer	-0.04	0.06	low	0.14	0.08	equal	-0.10	0.10	-1.84	252	.07	-0.12
	-0.04	0.06	low	-0.11	0.06	high	0.06	0.08	0.73	365	.47	0.04
	0.14	0.08	equal	-0.11	0.06	high	0.24	0.10	2.46	251	.01*	0.16

Note. *p*-value was adjusted with a Bonferroni correction ($\alpha = .05/3 = .0167$). [(a)]Effect size (*d*) was divided by two because Welch's t-test in the current study was a paired test (Dunlap et al., 1996). Effect size *d* = .20 is a small effect; *d* = .50 is a medium effect; and *d* = .80 is a large effect (Cohen, 1992, p. 157). ***p* < .01, **p* < .05.

Furthermore, *low* complexity tasks were found to be easier than *equal* (by 0.28 logits) and *high* complexity tasks (by 0.38 logits) graded in the *Answer Only* rating context (t[254] = -2.90, p = .00; t[365] = -4.69, p = .00, respectively), while no statistically different difficulty levels were found among *high*, *equal*, and *low* complexity level prompts in *Question and Answer* rating context, except between *equal* and *high* complexity prompts (t[251] = 2.46, p = .01), as shown in Table 4.26. The significance level with the rating context and task

complexity in Table 4.26 was adjusted with a Bonferroni correction ($\alpha = .05/3 = .0167$) for making three pairwise comparisons (Field, 2009).

Figure *4.9* graphically depicts the task difficulty difference of task complexity between *Answer Only* and *Question and Answer* rating contexts, as was displayed in Table 4.25. The graph shows that the difficulty of *high* complexity tasks plummeted from the *Answer Only* rating context to the *Question and Answer* rating context (statistically significant at $p < .01$), while that of *low* complexity tasks slightly increased (statistically significant at $p < .10$). The task difficulty of *equal* complexity tasks was not statistically different across the two rating contexts.

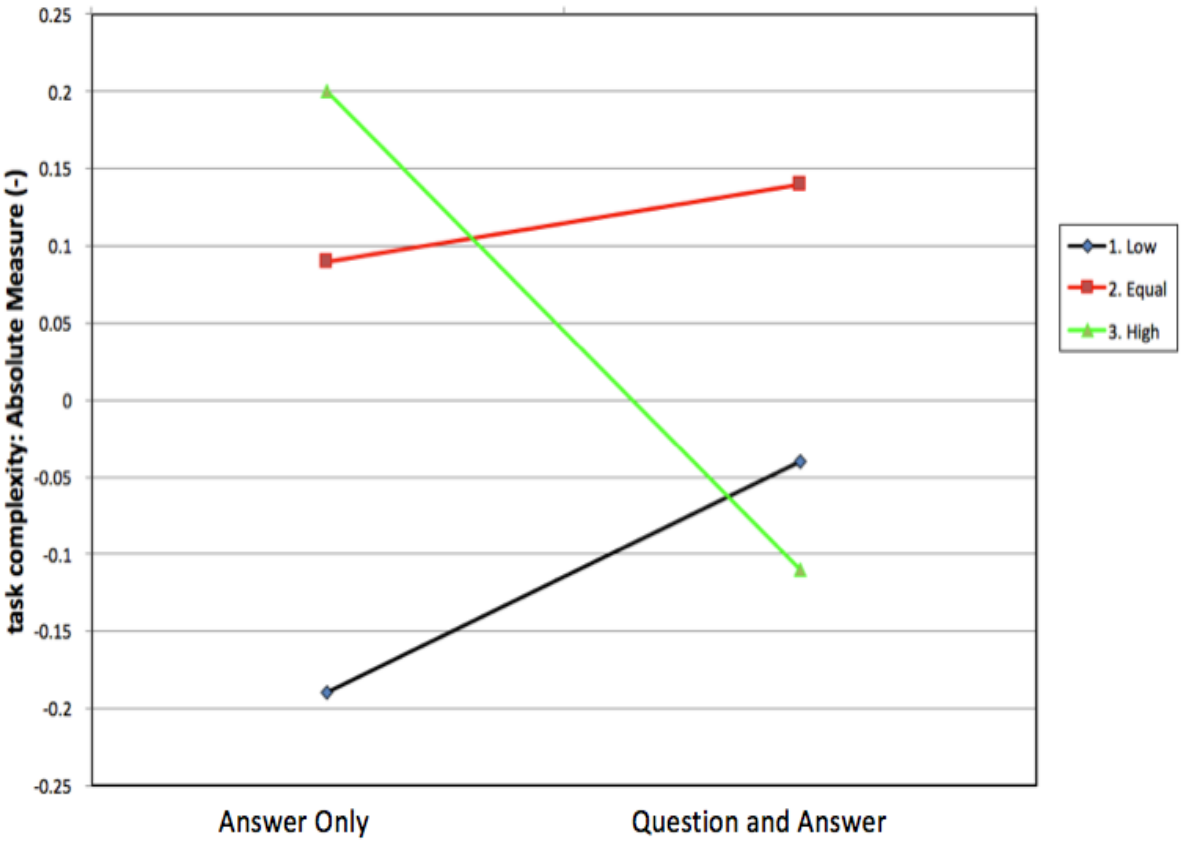


Figure 4.9. Comparison of average prompt difficulty in the *Answer Only* and *Question and Answer* rating contexts (SA group).

This different pattern of task difficulty change from *Answer Only* to *Question and Answer* rating contexts suggests that raters in the SA group might have changed their rating severity depending on their understanding of the task prompts for the following reasons. First, the insignificant change of task difficulty of the *equal* complexity tasks from the *Answer Only* to the *Question and Answer* rating contexts indicates that the *equal* complexity can be used as the reference group with which the change of task difficulty of *low* or *high* complexity tasks can be compared. Second, as the change of task difficulty from the *Answer Only* to *Question and Answer* rating contexts could be caused by both the effect of the rating context difference and the time interval between the two rating contexts (*Question and Answer* rating was conducted five days after *Answer Only* rating), the consistent task difficulty of *equal* complexity tasks in the two rating contexts ensures that the change of task difficulty of *high* or *low* complexity tasks could not be solely attributed to the bias caused by the rating time interval between the two rating contexts. Thus, it can be claimed that there was a task complexity interaction with the rating contexts, which supports the hypothesis that raters changed their rating severity depending on their understanding of the complexity level of the task prompts.

Ratings by Scale Non-Adjusting Group raters

***Main effects*.** The log-likelihood chi-square value of the MFRM model for raters who did not consider task complexity in their scoring was 1413.66 (df = 388, p = .00), indicating significant lack of global data fit (Eckes, 2015; Linacre, 2012). As was discussed with the models in earlier sections, this lack of global data fit does not necessarily mean that the model did not fit the data, because global fit statistics are sensitive to sample sizes (Eckes,

2015). The mean residual and the mean standardized residual were close to 0.00 and the standard deviation of the standardized residual was 1.00, thus suggesting successful estimation with the proposed model (Linacre, 2014). This Rasch model prediction of the variance of Rasch measures was 75.92%.

Table 4.27 shows the summary Rasch statistics of the PCM model for the raters in the Scale Non-Adjusting (SNA) group. The mean proficiency score of the test takers was 0.00 in logits, but the rater severity, task complexity, and rating context were centered at 0.00, as was done with the Scale Adjusting (SA) group. The infit mean-squares of the test taker, rater, task complexity, and rating context facets were 1.01, 1.01, 1.00, and 1.01, and their corresponding outfit mean-squares were 1.01, 1.01, 1.00, and 1.01. The infit and outfit mean-squares of the four facets were all within an acceptable range (0.5-1.5) (Linacre, 2014), indicating the model fit the data appropriately. The fixed chi-square values of the four facets were all statistically significant, which means that at least two levels of test takers, raters, task complexity, or rating contexts did not share the same value of test taker ability, rater severity, task complexity, and rating contexts, respectively. The separation ratio G and the strata index H of the test takers were 5.51 and 7.68, indicating that the test could statistically distinguish between more than six proficiency levels of test takers (Linacre, 2014). The strata index H for rater, task complexity, and rating context facets were 10.90, 3.75, and 2.86, indicating that there were nearly 10 statistically distinct classes of rater severity, more than three classes of task complexity, and at least two classes of rating context, respectively. The reliability of the separation for the test takers, raters, task complexity, and rating context were .97, .98, .87, and .78.

Table 4.27. *Rasch Measurement Summary Statistics (SNA Group)*

	Test takers	Raters[a]	Task Complexity[a]	Rating Context[a]
Measures				
Mean	0.00 (0.26)	0.00 (0.07)	0.00 (0.07)	0.00 (0.06)
RMSE	0.26	0.07	0.07	0.06
Adjusted (True) SD	1.45	0.56	0.19	0.11
Infit MS				
Mean	1.04	0.97	0.98	0.97
Outfit MS				
Mean	1.01	1.01	1.01	1.01
Fixed chi-square (model)	1046.5**	187.7**	22.5**	9.2**
df	39	2	2	1
Separation ratio (G)	5.51	7.93	2.57	1.89
Separation (strata) index (H)	7.68	10.90	3.75	2.86
Separation reliability (R)	.97	.98	.87	.78

Note: [a]Raters, task complexity, and rating context variables were centered at zero. $^{**}p < .01$.

Figure 4.10 presents the Wright map resulting from the partial credit analysis of the 40 test takers with three task complexity levels scored by raters in the SNA group. The first column in the Wright map provides a logit scale and the second column, test taker, shows the ability of test takers. The third column presents information about three raters in the SNA group. As shown in Table 4.28, the infit and outfit mean-square values of the raters were between 0.82 and 1.26, and they were all within the acceptable range (0.5-1.5) (Linacre, 2014), indicating that raters behaved as the Rasch model expected.

Table 4.28. *Rater Severity and Infit and Outfit Mean-squares (SNA Group)*

Raters	Rater Severity		Infit Mean-Square	Outfit Mean-Square
	Measures (logits)	SE		
R1	0.67	0.07	1.22	1.26
R7	0.03	0.07	0.82	0.91
R8	-0.70	0.07	0.89	0.85

The rating context for raters in the fourth column provides information about how raters judged the test taker audio files depending on their knowledge of task complexity. The rating scale difficulty levels (and standard errors) of the prompts when raters knew the questions (*Question and Answer*) and did not know the questions (*Answer Only*) were -0.12 (0.06) and 0.12 (0.05). The reported strata value was 2.86, with a reliability of .78, indicating that it is somewhat likely that the two rating contexts were statistically different in terms of their difficulty. As the difficulty level of the *Question and Answer* rating context is lower than that of the *Answer Only* context, raters would have given higher scores to test takers when scoring in the *Question and Answer* than *Answer Only* context.

The fifth column indicate the rating scale difficulty (and standard errors) of the prompts with *low*, *equal,* and *high* task complexity levels, which were -0.03 (0.06), -0.23 (0.09) and 0.26 (0.06), as shown in **Error! Reference source not found.**. In addition, the r eported strata value was 3.75 with a reliability of .87, indicating that at least two complexity levels were statistically different in terms of their difficulty.

Table 4.29. *Task Difficulty and Infit and Outfit Mean-Squares (SNA Group)*

Task Complexity	Difficulty		Infit Mean-Square	Outfit Mean-Square
	Measures (logits)	SE		
high	0.26	0.06	1.06	1.06
mid	-0.23	0.09	0.96	1.04
low	-0.03	0.06	0.90	0.93

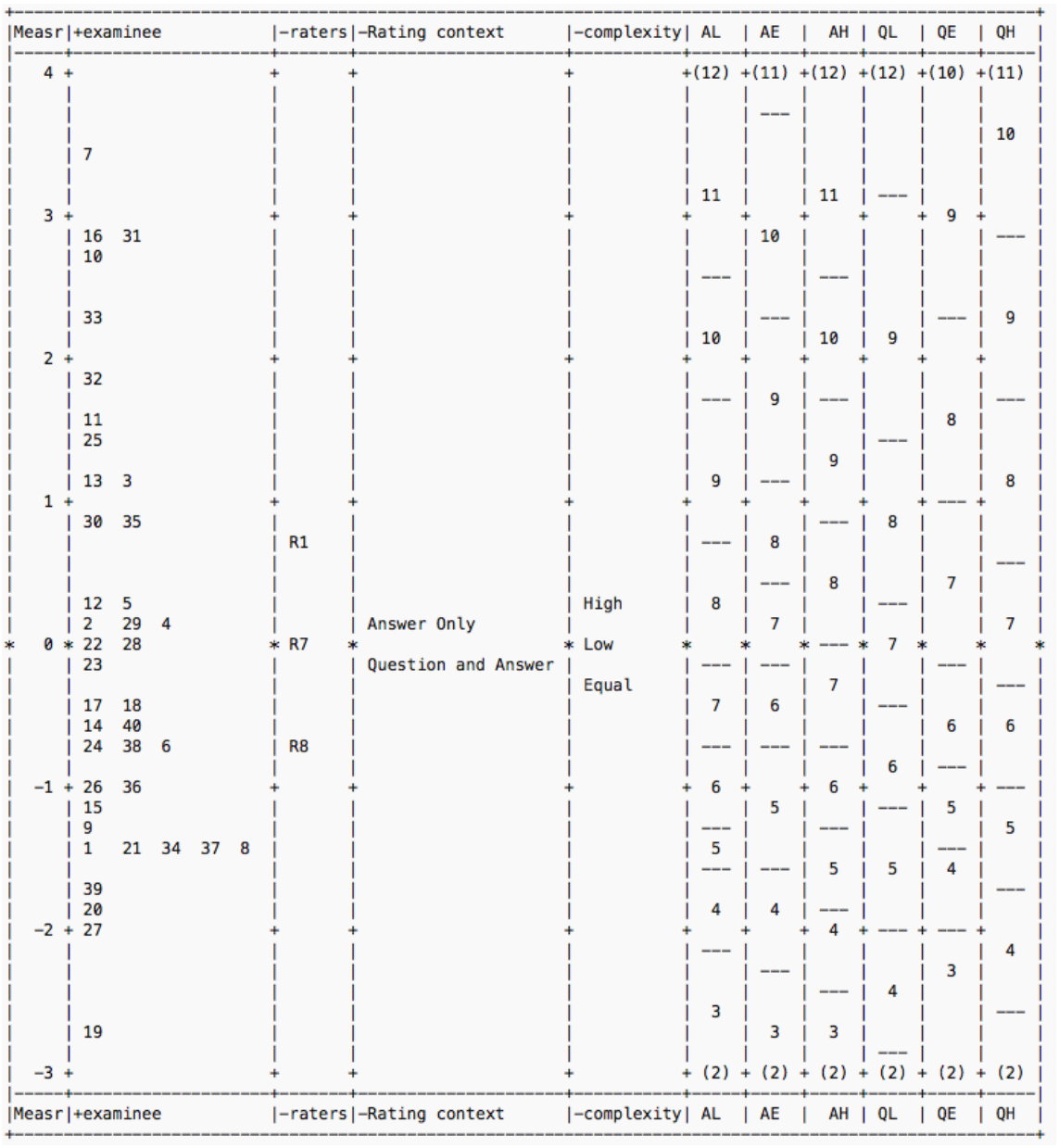


Figure 4.10. Wright map from raters who did not consider task complexity (SNA group)

Note. AL = low task complexity in an answer only context, AE = equal task complexity in an answer only context, AH = high task complexity in an answer only context, QL = low task complexity in a question and answer context, QE = equal task complexity in a question and answer context, QH = high task complexity in a question and answer context.

As was in the partial credit model for the raters in SA group, the last six columns provide information about the score categories with *Rasch-Andrich thresholds* for the rating scales of three task complexities in the two rating contexts.

Thresholds. Table 4.30 shows the average measures and standard errors of *the Rasch-Andrich thresholds* for the raters who indicated they did not consider task complexity. As shown in Figure *2.1* and Table 4.30, the *Rasch-Andrich thresholds* for *low*, *equal*, and *high* complexity prompts in the *Answer Only* rating context were not statistically different from those in the *Question and Answer* context considering the large standard error of each threshold. These large standard error estimates could be associated with the small sample size for each score category. In addition, as there were some missing observations in the higher score categories, such as 11 in QL, 11 and 12 in QE, an individual threshold comparison was not considered a valid method to examine raters' use of rating scales depending on task complexity in two different rating contexts.

Task complexity interaction with rating context. To estimate the size of any rating context effect (i.e., *Answer Only* and *Question and Answer* rating context) on the rating scale difficulty with the SNA group raters, the interaction of task complexity with rating contexts was analyzed. As was modeled with the SA group raters, rating context and task complexity were coded as dummy facets. Table 4.31 shows the pairwise bias report for the raters in the SNA group. *High* complexity tasks were 0.46 logits more difficult in the *Answer Only* rating context than those in the *Question and Answer* rating context (t[189] = 3.79, p = .00). In contrast, *equal* and *low* complexity tasks showed no statistical difference in their difficulty levels depending on their rating contexts (*equal* complexity: t[93] = 0.83, p=.41; *low* complexity: t[188] = 0.50, p = .62).

Table 4.30. *Rasch-Andrich Thresholds of the Rating Scale for Raters (SNA Group)*

	AL		AE		AH		QL		QE		QH	
Score	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SE	Threshold	SE
3	-2.01	0.56	-1.14	0.76	-2.28	0.52	-2.07	1.15	-2.02	0.76	-3.71	0.97
4	-2.52	0.45	-3.64	0.68	-2.23	0.42	-3.97	0.83	-2.05	0.61	-2.33	0.46
5	-0.79	0.38	-1.47	0.46	-2.19	0.36	-1.76	0.41	-0.90	0.51	-1.88	0.35
6	-1.72	0.36	-0.33	0.45	-0.85	0.32	-1.08	0.33	-1.24	0.47	-0.63	0.31
7	-1.16	0.32	0.21	0.49	-1.02	0.32	-0.62	0.30	-0.69	0.43	-0.45	0.31
8	0.11	0.32	-0.14	0.50	-0.03	0.33	0.36	0.31	1.37	0.45	0.15	0.33
9	0.27	0.34	1.06	0.50	0.90	0.39	1.02	0.35	1.88	0.54	2.21	0.42
10	2.17	0.43	2.43	0.63	1.79	0.46	3.07	0.45	3.64	1.03	2.31	0.49
11	2.23	0.49	3.01	1.09	2.33	0.52	-	-	-	-	4.34	0.84
12	3.43	0.64	-	-	3.59	0.76	5.05	0.99	-	-	-	-

Note. AL = low task complexity in an answer only context, AE = equal task complexity in an answer only context, AH = high task complexity in an answer only context, QL = low task complexity in a question and answer context, QE = equal task complexity in a question and answer context, QH = high task complexity in a question and answer context.

Table 4.31. *Pairwise Bias Report for Task Complexity with Rating Context (SNA Group)*

	Answer Only		*Question and Answer*				*Welch's t*			
Task Complexity	*Measure*	*SE*	*Measure*	*SE*	*Contrast*	*SE*	*t*	*df*	*p*	$d^{(a)}$
high	0.38	0.08	-0.08	0.09	0.46	0.12	3.79	189	0.00**	0.28
equal	-0.09	0.12	-0.24	0.13	0.15	0.18	0.83	93	0.41	0.09
low	-0.08	0.08	-0.14	0.10	0.06	0.06	0.50	188	0.62	0.04

Note. [(a)]Effect size (d) was divided by two because Welch's t-test in the current study was a paired test (Dunlap et al., 1996). Effect size $d = .20$ is a small effect; $d = .50$ is a medium effect; and $d = .80$ is a large effect (Cohen, 1992, p. 157). $^{**}p < .01$, $^{*}p < .05$.

Table 4.32 shows that the difficulty of *high* complexity task was statistically more difficult than that of *low* and *equal* complexity tasks only in the *Answer Only* rating context with the significance level adjusted with a Bonferroni correction ($\alpha = .05/3 = .0167$) (vs. *low*

complexity: *t*[190] = -3.99, *p* = .00; vs. *equal* complexity: *t*[106] = -3.22, *p* = .00). On the other hand, the difficulty of the *high* complexity task was not statistically different from that of *low* or *equal* complexity tasks in the *Question and Answer* rating context. Figure *4.11* visually presents the change of task difficulty from *Answer Only* rating context to *Question and Answer* rating context. As was shown in Table 4.31, the difficulty of *high* complexity tasks greatly decreased from *Answer Only* context to *Question and Answer* rating context, while *low* and *equal* complexity tasks were slightly decreased with no statistical significance, indicating that even raters in the SNA group may have changed their rating severity once they got to know the complexity level of the prompts.

Table 4.32. *Pairwise Bias Report for Rating Context with Task Complexity (SNA Group)*

	Task 1			*Task 2*					*Welch's t*			
Rating Context	*Measure*	*SE*	*Complexity*	*Measure*	*SE*	*Complexity*	*Contrast*	*SE*	*t*	*df*	*p*	$d^{(a)}$
Answer Only	-0.08	0.08	low	-0.09	0.12	equal	0.01	0.15	0.10	106	0.92	0.01
	-0.08	0.08	low	0.38	0.08	high	-0.46	0.12	-3.99	190	0.00**	-0.29
	-0.09	0.12	equal	0.38	0.08	high	-0.47	0.15	-3.22	106	0.00**	-0.32
Question and Answer	-0.14	0.10	low	-0.24	0.13	equal	0.10	0.16	-0.60	113	0.55	-0.6
	-0.14	0.10	low	-0.08	0.09	high	-0.06	0.13	-0.41	189	0.68	-0.03
	-0.24	0.13	equal	-0.08	0.09	high	-0.15	0.16	-0.96	111	0.33	-0.09

Note. *p*-value was adjusted with a Bonferroni correction (α = .05/3 = .0167). [(a)]Effect size (*d*) was divided by two because Welch's t-test in the current study was a paired test (Dunlap et al., 1996). Effect size *d* = .20 is a small effect; *d* = .50 is a medium effect; and *d* = .80 is a large effect (Cohen, 1992, p. 157). ***p* < .01, **p* < .05.

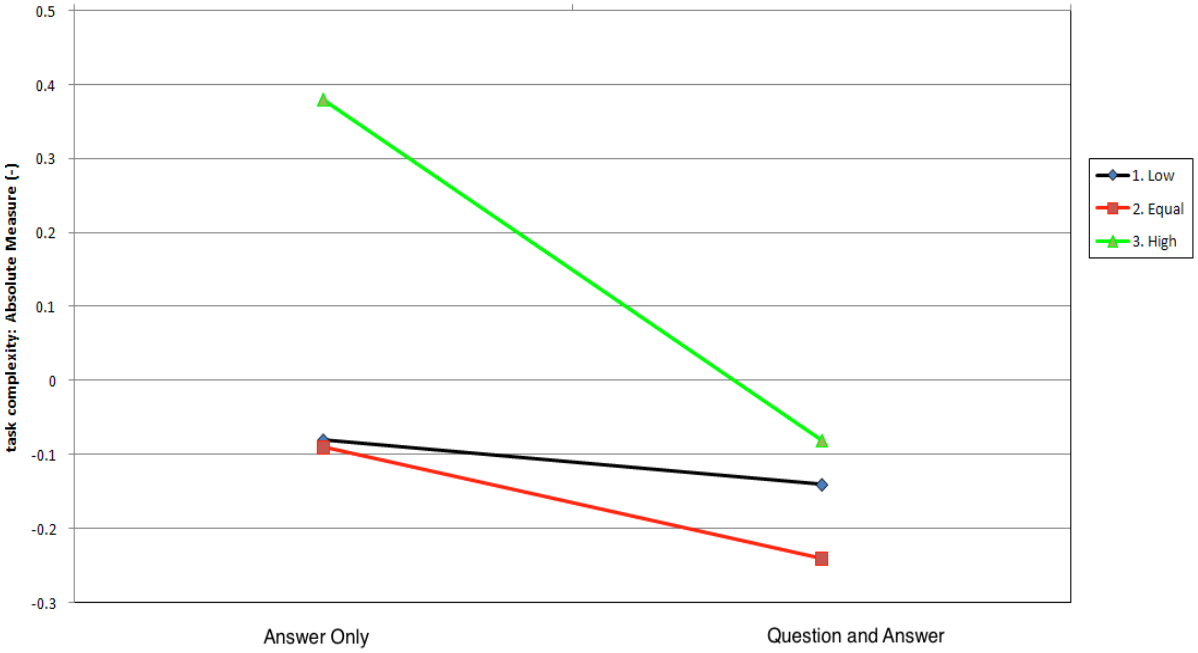


Figure 4.11. Comparison of average prompt difficulty in the *Answer Only* and *Question and* Answer rating contexts (SNA group).

4.3.2. Section Summary

This section presented the results of two partial credit models of MFRM with the data in an experimental condition, which controlled the adaptiveness of the OECT test. The first MFRM model was used to investigate the change of rater severity by task complexity for the six raters in the Scale Adjusting group. The task complexity interaction with a rating context analysis using Welch's *t*-test showed that the difficulty of *high* complexity tasks decreased while that of *low* complexity tasks slightly increased when raters thought that it was difficult for test takers. The second MFRM model was used for the three raters in the Scale Non-Adjusting group. Even though these raters claimed that they did not take task complexity into account when scoring test takers' performance, the task complexity interaction with rating context analysis revealed that the difficulty of *high* complexity tasks decreased after raters got to know the task complexity. This change of task difficulty with raters both in the Scale

Adjusting and Non-Adjusting groups suggests that raters in an adaptive performance-based oral communication might have changed their rating severity depending on their understanding of the task complexity.

4.4. Qualitative Analysis of Rating Scale Use (Sub-Study 2)

The result of the quantitative analysis of raters' rating scale use generally supports the hypothesis that raters adjusted their rating severity depending on their perception of task complexity. However, the quantitative analysis did not provide information about what caused raters to adjust their rating severity in relation to task complexity. A qualitative analysis was needed to fill in the logical gaps between the findings of the quantitative analysis and the hypothesis of this study, which is that raters adjusted their rating severity depending on task complexity. The raters' verbal reports were first examined to investigate how raters incorporated task complexity into their interpretation of rating descriptors in the scoring rubric--*functional competency* (FC), *fluency* (PF), *lexico-grammar* (LG), and *pronunciation* (PR)--with *task complexity* (TC), during the experimental rating session. Interview data were also examined to more explicitly understand the effect of raters' perception of task complexity on their rating severity.

4.4.1. Analysis of Verbal Reports

After two rating sessions of 80 audio clips with and without task prompts in the experimental rating session, nine raters were asked to record verbal reports while scoring eight audio clips from the 80 audio clips in an experimental study (see Appendix C for verbal report procedures). The recorded retrospective verbal reports were analyzed to understand

how raters considered task complexity while evaluating the proficiency level of test takers' performance; however, raters were not explicitly guided to discuss task complexity during the verbal report session. The following two research questions were addressed in the verbal report analysis.

RQ 2-2. Which evaluation categories do raters attend to for task complexity while scoring oral communication audio clips?

RQ 2-3. How do raters apply task complexity to their interpretation of evaluation criteria in terms of rating severity?

Transcription and coding of verbal reports

An orthographic transcription of the verbal reports was carried out, but the transcription did not include pauses or hesitations, because this study focused not on the raters' language use, but on their rating process. The word count of about 50 minutes' verbal reports ranged from 1,547 to 2,771. The transcription of Rater 1's verbal report (approximately 10% of the total verbal report) was coded by two coders, whose percent agreement was 0.75 and *Krippendorff's alpha* (Krippendorff, 1970, 2004a) was 0.67, indicating there was a moderate inter-coder agreement (Krippendorff, 2004b). This moderate level of reliability in the coding was deemed acceptable for this study, but results should be considered in light of this somewhat less than desirable consistency. The transcriptions of Raters 2 to 9 were coded solely by the first coder. As shown in Table 4.33, the number of codes per each rater ranged from 73 to 124, and *functional competency* (n = 168) and *fluency* (n = 177) were the most frequently used codes by the raters. The *task complexity* (n = 118) was also somewhat frequently mentioned in the verbal reports.

Co-occurring coding category with task complexity

Five coding categories in Table 4.33, except task complexity and interviewer's behavior, were examined to inspect the evaluation categories to which raters attend together with task complexity. As shown in Table 4.34, *functional competency* [*n* = 22 for raters in the Scale Adjusting (SA) group; *n* = 5 for Scale Non-Adjusting (SNA) group] and *fluency* (*n* = 17 for raters in SA group; *n* = 7 for raters in the SNA group) most frequently co-occurred with task complexity. This highly frequent co-occurrence of *functional competency* and *fluency* could be due to their close relationship to task complexity. On the other hand, *lexico-grammar* (*n* = 9) and *pronunciation* (*n* = 4) least frequently co-occurred with task complexity. Even after considering the total frequency of the evaluation categories, the relative frequency of the co-occurrence of *functional competency* and *fluency* with task complexity was higher than *lexico-grammar* and *pronunciation*.

Table 4.33. *Word Count and Coding for Nine Raters' Verbal Report.*

Rater	Word Count	Coding Category							
		GP	FC	PF	LG	PR	TC	IB	Total
R1	2,229	7	22	14	4	10	13	2	72
R2	2,248	18	23	28	17	17	21	0	124
R3	2,031	3	28	14	9	7	9	7	77
R4	1,886	7	17	12	10	15	14	3	78
R5	2,279	2	14	16	15	13	16	6	82
R6	1,547	13	16	23	15	15	12	1	95
R7	2,455	12	17	23	20	20	10	0	102
R8	1,778	5	10	17	15	23	10	0	80
R9	2,771	11	21	30	13	27	12	0	114
Total	19,224	78	168	177	118	148	117	19	824

Note. GP = General proficiency not included in FC, PF, LG, and P; FC = functional competency; PF = pace and fluency; LG = lexico-grammar; PR = pronunciation, TC = task complexity; IB = interviewer's behavior.

Table 4.34. *Co-occurring Evaluation Categories with Task Complexity.*

Group	Evaluation Category	Co-occurrence	Total Frequency	Co-occurrence *100 / Total Frequency (%)
Scale Adjusting Raters (n=6)	General Proficiency (GP)	8	54	14.81
	Functional Competency (FC)	22	119	18.49
	Fluency (PF)	17	123	13.82
	Lexico-grammar (LG)	8	79	10.13
	Pronunciation (PR)	3	94	3.19
	Total	58	469	12.37
Scale Non-Adjusting Raters (n=3)	General Proficiency (GP)	6	24	25.00
	Functional Competency (FC)	5	49	10.20
	Fluency (PF)	7	54	12.96
	Lexico-grammar (LG)	1	39	2.56
	Pronunciation (PR)	1	53	1.89
	Total	20	219	9.13

Scale adjusting group. The six raters in the SA group (R2-6 and R9), who explicitly expressed that they did consider task complexity in interviews, verbalized how they interpreted task complexity along with the rating scales while they were scoring the eight audio clips. The verbal reports of the SA group showed that the raters considered task complexity while scoring test takers' audio clips. Among the raters in the SA group, Rater 2, in the following quote, considered task complexity in relation to *functional competency*:

> **Excerpt 4.1. Rater 2 / Scale-Adjusting Raters Group / Rating with Questions / Functional Competency**
> […] it's just repeating things again and again. And, that's that didn't trouble me a lot. It's because I think it is because of the question, the difficulty of the question [*TC*] because […] I don't think he could think any more idea about any more reasons about that topic [*FC*].

Rater 2 could have focused on the test taker's performance without considering task complexity, but he interpreted the evidence of *functional competency* along with task complexity by mentioning that the test taker's repetition was "because of […] the difficulty of the question". This consideration of task complexity could be due to the rater's understanding of descriptors in the *functional competency* category which contained topical

points, such as “familiar topics and handle unsophisticated tasks.” Among four evaluation categories in the scoring rubric, *functional competency* was the only category that contained the topical elements. The descriptors in the *functional competency* category could be among the factors that influenced the high frequency of *functional competency* with task complexity.

The following quote by Rater 2 supports the hypothesis that the rater considered task complexity even with the scoring categories (e.g., fluency) that did not contain task complexity in their descriptors. When Rater 2 evaluated the fluency level of test takers who had a *high* complexity prompt, he reported that the “silence” by the test takers before the first response to the prompt could be accepted as an excuse for poor fluency because the test taker needed to “think about” the question and response.

> **Excerpt 4.2. Rater 2 / Scale-Adjusting Raters Group / Rating with Questions / Fluency**
> I like her idea this part and the first three seconds that when there was silence [*PF*] didn't really trouble me. Although she could use some get fillers there. But I just I'm okay with it because the because of the question [*TC*]. She was just thinking about her response.

Rater 2, as shown in the following quote, seems to have had an understanding of the relationship between task complexity and linguistic performance. According to studies associated with Robinson’s (2001, 2011b) Cognition Hypothesis, more complex prompts are believed to trigger higher levels of vocabulary and grammar use (Kormos, 2011). Rater 2 reported that he did not expect test takers “to use a lot of complicated vocabulary” with an easy topic while evaluating the vocabulary and grammar features of the performance. The rater’s low expectation of vocabulary matches with findings of the studies about the effect of task complexity on vocabulary use (Kormos, 2011). If Rater 2 did not understand the effect

of task complexity on test takers' linguistic outputs, his interpretation of test takers' performance could have been different.

> **Excerpt 4.3. Rater 2 / Scale-Adjusting Raters Group / Rating with Questions / Lexico-grammar**
> Although the topic is easy [*TC*], she just her grammar and the choice of vocabulary was effective. The topic is easy, [so] you cannot we don't expect them to use a lot of complicated vocabulary [*LG*], but she didn't really use it in a very simple way either [*LG*].

With regard to the effect of task complexity on raters' scoring practice, the following quotes by Rater 3, Rater 4, and Rater 5 show more direct evidence in support of the hypothesis that raters adjusted their rating severity depending on the complexity of the prompts. As described in Excerpt 4.4, Rater 3 mentioned that he took "the difficulty of the topic" into account when judging a test taker's performance score. It seems that Rater 3 did not penalize the test takers' lack of elaboration and fluency, because the rater was aware that the complexity level of the topic was quite high.

> **Excerpt 4.4. Rater 3 / Scale-Adjusting Raters Group / Rating with Questions / Fluency**
> I think she's repeating her idea. She was not quite able to elaborate more on her ideas [*FC*]. For this topic, it's hard. Even though she became a little choppy [*PF*], but taking into consideration the difficulty of the topic [*TC*], I think [the score is] eight.

Rater 4 added more specific comments to his rating process by saying the "question really determines […] the way I score him." In addition, Rater 4 used the task type, or *abstract* question, as one of his rating criteria. These statements seem to provide evidence to the hypothesis that Rater 4 has changed his rating severity depending on task complexity.

Excerpt 4.5. Rater 4 / Scale-Adjusting Raters Group / Rating with Questions / Functional competency

The question really determines the way, perhaps influence the way I score him, because this question is quite tough, and somehow, she can answer the question. Yeah, she can explain. And this is quite abstract topic [*TC*], even though her pronunciation is not that clear. But then it's much better than the previous one. And then, so I gave her *adequate* because she can manage to answer this abstract question [*FC*].

Rater 5 also reported, in the following quote, that the test taker could have developed the topic more thoroughly because it was "not very difficult." Rater 5's expectation from the test taker could differ from Rater 2, who reported in Excerpt 4.2 that simpler topics generally elicit simpler vocabulary and sentence structures; Rater 5 expected that test takers with simpler topics could produce better *functional competency* while Rater 2 believed that test takers could not use complicated vocabulary and grammar structures with easier prompts. Thus, the interpretation of simple expressions and simple topic development by Rater 2 and Rater 5 could be varied, because their expectations of the test takers' linguistic performance with *low* complex prompts are different.

Excerpt 4.6. Rater 5 / Scale-Adjusting Raters Group / Rating with Questions / Functional Competency

[…] the topic is not very difficult [*TC*]. It's most descriptive, and but he's just using very, very similar simple expressions […] these are very, um, simple structures for such a topic. I would expect more more topic development for such a relative was simple topic [*FC*]. Yeah, that's why I would give him six […]

Scale non-adjusting group. The three raters in the SNA group (R1, R7, and R8), who explicitly expressed that they did not consider task complexity in interviews, also verbalized how they interpreted task complexity along with other rating categories. As shown in Table 4.34, there were as many co-occurring codes with task complexity in the SNA group as in the SA group. In the following quote, Rater 1 commented on the effect of task complexity on test

takers' performance, but she used a hedging expression "I don't know if question has an effect on them" to describe that she did not directly consider task complexity in her rating process. It seems that the rater successfully analyzed the test taker's linguistic performance; however, the rater was not sure how she incorporated the task difficulty into her scoring when test takers were given *low* complex prompts.

> **Excerpt 4.7. Rater 1 / Scale Non-Adjusting Raters Group / Rating with Questions / General Proficiency**
> So she was talking really smooth. She was completely able to answer the question. I don't know if question has an effect on them, because it is one of the easy questions that you can compare with your own life. But she was doing it pretty well I guess.

On the other hand, Rater 7, in the following quote, reports that he was aware of the task complexity and its effect on the test taker's performance. Rater 7 explicitly reports that he did not give a lower score because of the "difficulty of the question," which is somewhat contradictory to his answers to the interview questions (in the following section) that he did not consider task complexity for scoring. The rater, however, was slightly hesitant to have more confidence in his perception of the effect of task complexity by saying the hedging expression "probably, but still, I can't say it is eleven. That's why I went with ten."

> **Excerpt 4.8. Rater 7 / Scale Non-Adjusting Raters Group / Rating with Questions / General Proficiency**
> The communication probably is not really effective, but that could be because of the question itself. The nature of the question. Question is a little bit difficult. It's about cultural sayings or cultural aspects of language, which might not be familiar to the test taker. The only reason I'm not saying lower than that for the communications effectiveness is the difficulty of the question. Probably, but still, I can't say it is eleven. That's why I went with ten, like a previous one.

Rater 8 also reports, in the following quote, that he was aware of task complexity while scoring test takers' performance. The rater, however, did not directly comment on whether he adjusted his rating scale. He instead used the hypothetical statement "if she got a

different question, she will get better scores" not to make any conclusive statement about the effect of task complexity on his rating severity.

> **Excerpt 4.9. Rater 8 / Scale Non-Adjusting Raters Group / Rating with Questions / General Proficiency**
> I will give her Eight. There was not good enough data to evaluate her speech in English, but the question was kind of fairly difficult to answer. She might get some confusion. I think when she got this answer so but she tried to say something, and so it's kind of not clear what she sort of the definition of rich and heavy money. But so, I think if she got a different question, she will get better scores.

Findings of verbal reports analysis

Both groups of raters indicated that they considered task complexity when rating test taker responses. In general, raters in the SNA group had a relatively smaller number of co-occurring codes with task complexity than those in the SA group. Unlike the original hypothesis and interview results that raters in the SNA group would not take task complexity into consideration while scoring, raters in the SNA group also reported co-occurring evaluation criteria with task complexity. These verbal reports by the raters in the SNA group, however, were somewhat different from those by the raters in the SA group in that raters in the SNA group used more hedging expressions, such as "I don't know if" or "probably" to soften their description of the direct relationship between task complexity and other evaluation criteria.

4.4.2. Analysis of Interview Questions

After completing verbal reports with eight audio clips, raters were interviewed about how they completed the questionnaire in Appendix D to investigate their rating preference and the effect of task complexity on their rating severity. This made it possible to more directly ask raters about their rating processes. Raters' responses to the questionnaire and

their interview reports were compared with the analysis of the retrospective verbal reports in the previous section to answer the following research question:

RQ 2-3. How do raters apply task complexity to their interpretation of evaluation criteria in terms of rating severity?

Relative importance of evaluation criteria

As shown in Table 4.35, when raters were asked to order the importance of evaluation categories in their scoring, most raters in the SA group responded that *fluency* and *pronunciation* were the two most important categories. Among the two important categories, *fluency*, instead of *pronunciation*, was considered as the main evaluation category to be of concern in the current study; this was because the quality of *pronunciation* generally would not vary depending on the task complexity. In addition, as *fluency* was one of the most frequently coded scoring criteria with task complexity, as shown in Table 4.34, the interaction between *fluency* and task complexity need to be further investigated to explain the score variance due to the effect of task complexity on raters' scoring.

Table 4.35. *Relative Importance of Evaluation Categories*

Group	Rater	Functional Competency	Fluency	Lexico-grammar	Pronunciation
Scale Adjusting Raters	R2	1	2	4	3
	R3	3	1	4	2
	R4	4	3	2	1
	R5	3	1	4	2
	R6	3	1	4	2
	R9	4	2	3	1
	Mean	3.00	1.67	3.50	1.83
Scale Non-Adjusting Raters	R1	3	2	4	1
	R7	1	3	4	2
	R8	1	3	2	4
	Mean	1.67	2.67	3.33	2.33

Note. 1-4 indicate the order of importance from the most important (1) to the least important (4) categories.

In follow-up interviews, each rater explained their rationale for ordering the importance of each of the evaluation categories. Rater 2 chose *functional competency* as the most important category, which was different from other raters in the SA group. Rater 2 understood that *functional competency* would indicate how successful the communication was.

> **Excerpt 4.10. Rater 2 / Scale Adjusting Group / Rating Category**
> Functional competency is more important than others to me. To me in real life, we do this like, I mean, we're looking at how successful you are exchanging the ideas. […] So it's very, very important than having a good pronunciation but couldn't express your ideas. That's the problem. So having a functional competency but expressing writers with a very bad pronunciation, that I would say that's not very problematic as long as the person you were talking to understands you. That's Okay.

On the other hand, Rater 6, in the following quote, emphasized the importance of *fluency* and *pronunciation* by explaining that communication would fail if test takers had poor *fluency* or *pronunciation*.

> **Excerpt 4.11. Rater 6 / Scale Adjusting Group / Rating Category**
> […] I considered the most is fluency and pronunciation. And then maybe functional competency and then vocabulary and grammar. […] Um, so especially for the pronunciation, If the speech is too accented. I still went for the lower grade, because I couldn't understand it. I think that should go before other categories like functional competence and vocabulary.

Both Rater 2 and Rater 6 situated communication as the baseline for the evaluation of test takers' performance, but their primary elements for successful communication were different, thus leading to different application of evaluation criteria. Rater 2 seems to believe that understandable *pronunciation* is sufficient for good communication, while Rater 6 seems to believe that good *pronunciation* improves the quality of communication.

Another interesting interview point from the SA group is Rater 4's rationale for his decision about the importance of evaluation criteria. Rater 4 reported, in the following quote, that he chose *functional competency* as the least important category not because *functional competency* was not important for successful communication, but because the category could be easily affected by task complexity. As the descriptors in the *functional competency* category were quite dependent on the topic, raters who considered task complexity in their rating might not have put much emphasis on *functional competency* to minimize unfair rating that was believed to be caused by task complexity.

> **Excerpt 4.12. Rater 4 / Scale Adjusting Group / Rating Category**
> […] Pronunciation is the most important thing in my opinion, whether I can understand this, what he's saying, what they're saying or not. […] Because why I consider this [functional competency] number four because some of the questions are very tough and it's not fair for them to. If you have, oh, very, very weak student and then you give them tough questions by very abstract one, then they will suffer […]

In contrast, Rater 7 and Rater 8 in the SNA group indicated that *functional competency* along with *pronunciation* were the most important evaluation categories. Even though these two raters did not provide a rationale for their ordering of the relative importance of evaluation criteria, this preference partially explains why these raters answered that they did not consider task complexity in their rating. As was discussed with the SA group, if raters considered *functional competency* as the important evaluation criterion, they must have had the confidence that they would not be biased in their interpretation of the descriptors in the *functional competency* category when the complexity of task was different. Unlike other raters in the SNA group, Rater 1 reported that she did not put much emphasis on *functional competency* not because it was not important, but because the duration of the interview audio clip was too short to evaluate *functional competency*.

Excerpt 4.13. Rater 1 / Scale Non-Adjusting Group / Rating Category
[…] I will start with pronunciation, although I feel bad for doing that, but I felt like words of Chinese students taking that I had a really hard time relating their pronunciations. Although I know that I shouldn't do that. I went with pronunciation, pace and fluency. […] Functional competence is three and vocabulary and grammar four. Functional competency I will say is too big for such a task. Sometimes the duration is so short that functional competences sound more like big to me. So I was not sure how I should relate this one minute question or one minute answer, like concrete topics, abstract topics, there is not much detailed information to evaluate their speech based on that one […]

Consideration of task complexity in scoring

Raters were asked to respond to three interview questions about how much they considered task complexity in their rating. The first question was about whether raters applied the same evaluation criteria for holistic scoring when test takers had to deal with prompts of different complexity levels. As shown in Table 4.36, the majority of raters in the SA group chose either *disagree* or *neither agree nor disagree*, while all raters in the SNA group chose *agree*. Rater 4 and Rater 5 disagreed with the statement that they applied the same evaluation criteria for the holistic scoring with different task complexity. Rater 4 explicitly reported in his interview that "if the level of difficulty is very high, then I tend to be more lenient in the score because this is not that fair." Rater 5 responded that she became "more flexible with the functional competency" when the level of task complexity was different. Even though Rater 2 and Rater 6 marked *strongly agree* and *agree*, they also reported they considered the type of questions in their scoring. Rater 2 said "the type of question [was] very important to me," and Rater 6 mentioned that she considered task complexity when she "debate[d] between two scores."

Table 4.36. *Raters' Consideration of Task Complexity in Scoring*

Group	Rater	Q1: Apply Same Evaluation Criteria for Holistic Scoring	Q2: Only Focus on Linguistic Outputs	Q3: Consideration of Task Complexity
Scale Adjusting	R2	(5) *(a)*	2	5
	R3	3	3	4
	R4	2	2	4
	R5	2	2	4
	R6	(4) *(b)*	2	5
	R9	2	2	4
	Mean	2.25	2.17	4.33
Scale Non-Adjusting	R1	4	4	2
	R7	4	4	2
	R8	4	2	2
	Mean	4.00	3.33	2.00

Note. 1: strongly disagree, 2: disagree, 3: neither agree nor disagree, 4: agree, and 5: strongly agree. *(a)* and *(b)* indicate raters' possible misunderstanding of the question during the interview, and these numbers were excluded for the mean value.

For the second question about raters' focus on test takers' linguistic outputs, raters in the SA group mostly chose *disagree*, indicating that they not only focused on linguistic outputs, but also considered task complexity when judging test takers' performance. On the other hand, raters in the SNA group mostly chose *agree*, indicating that they tried to focus only on the test takers' linguistic performance without considering task complexity during their scoring.

For the third question about raters' consideration of task complexity, used as the criterion to divide raters into the SA and SNA groups in this study, raters in the SA group responded that they *agreed* or *strongly agreed* with the statement that they considered task complexity when they judged test takers' performance. Raters did not elaborate much on the responses to this question. Rater 3 and Rater 5 briefly answered that they considered task complexity, but they could not explain how they incorporated task complexity into their scoring habits. Rater 3 mentioned "I don't know how much my rating reflected my thought,"

and Rater 5 reported "I subconsciously do […] if the task is very difficult then I revisit my criteria."

In contrast, raters in the SNA group responded that they *disagreed* with the statement that "I take the complexity (or difficulty) into consideration when I grade a test taker's performance". Rater 1 stated that she "felt bad for some students who had difficult questions […] but focused on the rubric rather than question." Even though Rater 1 was categorized as a rater in the SNA group, her comment indirectly suggests that she was at least thinking of task complexity while rating test takers' performance. Rater 7 in the SNA group also stated, in the following quote, that task complexity would affect test takers' performance, but he kept his position in the SNA group by saying "every low speed does not mean lack of fluency." By expanding the generally accepted mechanical definition of *fluency*, such as *Speech Rate*, *Mean Length of Utterance*, and *Phonation-Time Ratio* discussed in the linguistic analysis section, to a more flexible definition including dysfluency measures, such as fillers, Rater 7 seemed to differentially interpret test takers' performance at different task complexity levels while believing that he consistently applied the same rating criteria. This discrepancy between raters' actual rating behaviors and their rating beliefs could have led to the quantitative results, which are similar to those by the raters in the SA group, that the difficulty of *high* complexity prompts has changed from *Answer Only* to *Question and Answer* rating contexts.

> **Excerpt 4.14. Rater 7 / Scale Non-Adjusting Group / Rating Category**
> […]. I mean every slow speed does not for me does not mean lack of fluency. Because native speakers may also encounter questions that I can't answer. So they know how to fill the gaps and use the fillers strategies that compensate for the lack of fluency. So I'm going to say disagree.

Consideration of interviewer's performance

In the last part of the interview about how raters completed the questionnaire, raters were asked to respond to questions about their general consideration of interviewers' performance in the oral communication assessment. As the interaction between interviewers and test takers was minimized in the audio clips for the current study, raters' responses to the interview question would only represent their general perception of interviewers' performance irrespective of the ratings in the current study.

Table 4.37. *Raters' Perception and Consideration of Interviewers' Performance*

Group	Rater	Inappropriate Topic Choice	Unhelpful Questions	Closed Follow-up Questions	Over-domination
Scale Adjusting Raters	R2	5	4	4	4
	R3	2	4	2	4
	R4	3	2	2	2
	R5	4	3	4	5
	R6	3	3	3	4
	R9	4	2	2	4
	Mean	3.50	3.00	2.83	3.83
Scale Non-Adjusting Raters	R1	2	2	2	2
	R7	5	5	3	5
	R8	4	4	3	4
	Mean	3.67	3.67	2.67	3.67

Note. 1: strongly disagree, 2: disagree, 3: neither agree nor disagree, 4: agree, and 5: strongly agree.

As shown in Table 4.37, there were no saliently different response patterns between raters in the SA group and those in the SNA groups. Raters generally agreed with the statement that they considered interviewers' inappropriate topic choice and over-domination of the conversation during the oral interview test. About interviewers' unhelpful questions or closed follow-up questions, raters generally did not lean toward either *strongly agree* or *disagree*. This finding would partially justify the presumption that the results of the current study with task complexity are less influenced by the potential interviewer-related biases, such as

interviewers' competency (Morton et al., 1997) and helpfulness (Chartrand & Bargh, 1999; Clark, 2002; Wilson & Wilson, 2005).

Findings of interview analysis

In response to the question about the relative importance of each scoring category, raters in the SA group responded that *fluency* and *pronunciation* were the two most important categories while those in the SNA group indicated that *functional competency* along with *pronunciation* were the most important evaluation categories. Unlike those in the SNA group, raters in the SA group generally de-emphasized the importance of *functional competency*, because they believed that *functional competency* could be easily affected by task complexity, leading to unfair scoring. In response to the question about raters' consideration of task complexity in scoring, raters in the SA group generally agreed that they did consider task complexity while raters in the SNA group did not. Even though raters in the SNA responded that they did not consider task complexity in scoring, they also commented on task complexity several times in the follow-up interview. For the question about raters' consideration of interviewers' performance in scoring, there were no response differences between raters in the SA group and those in the SNA group, which would eliminate potential biases caused by interlocutors when comparing the behaviors of the SA and SNA groups in the current study.

4.4.3. Section Summary

This section presented the results of verbal reports and interview analyses that examined how raters considered task complexity while scoring test takers' responses. For the

verbal report analysis, raters' verbal reports were first coded based on the evaluation categories in the scoring rubric. Raters in the SA group were found to have declared use of more task complexity-related criteria than those in the SNA group. Even though raters in the SNA group explicitly stated in an interview that they did not consider task complexity during scoring, they also made several comments in their verbal reports about their consideration of task complexity along with other scoring categories. The results of verbal reports and interview analyses support findings in the quantitative analyses, where both groups of raters were found to have changed their rating severity depending on their knowledge of task complexity.

CHAPTER 5. DISCUSSION AND CONCLUSION

The purpose of the present study was to investigate raters' scoring behavior when they knew the task complexity before scoring test takers' performance in an adaptive performance-based oral communication test. For this purpose, operational data from an oral communication test for international teaching assistants (ITAs) and samples of operational data in an experimental condition were examined by conducting two sub-studies based on a sequential exploratory mixed-methods design (Creswell, 2014). With the operational data, the adaptiveness of the ITA test was checked and raters' behaviors in the operational test setting were quantitatively investigated. As is usually the case when conducting research with operational data (Tarone, 1998), there were some limitations in that the findings of the study could include large unexplained variance.

To overcome the limitations in the study with the operational data, a more controlled experimental study with samples of audio clips from the operational data was conducted. The rater behavior in the experimental study was analyzed both quantitatively and qualitatively with six raters in a Scale Adjusting (SA) group (raters who indicated that they took task complexity into account when providing ratings) and three raters in a Scale Non-Adjusting (SNA) group (raters who indicated that they did not take task complexity into account when providing ratings). The findings of the experimental study supported the hypothesis that raters generally changed their rating severity depending on their understanding of task complexity. The qualitative analysis of verbal reports and interviews using an interview protocol was conducted to understand how and why the raters recalibrated their rating scales. The qualitative analysis provided further evidence that raters generally considered task complexity while rating test takers' oral communication performance. Raters from both the

SA and SNA groups stated in their verbal reports that they adjusted their rating severity, even though raters from the SNA group responded in an interview questionnaire that they did not consider task complexity while scoring.

In this concluding chapter, raters' consideration of task complexity in the two quantitative studies, along with the findings in the qualitative analysis, is summarized and discussed. Discussion of theoretical, methodological, and practical implications are also presented. Finally, limitations of the current study and directions for future research are suggested.

5.1. Summary and Discussion of the Main Findings

The main findings of the study are summarized and discussed in response to each of the research questions. The first main research question was examined with the operational OECT data scored by 24 experienced raters, and the second main question was examined with scoring, think-aloud, and interview data by nine newly trained raters in a controlled experimental context.

5.1.1. Summary of the Main Findings

Rating scale use with the operational data (RQ 1)

The first main research question examined how experienced raters of the OECT adjusted their score assignment depending on the complexity level of prompts in an adaptive performance-based oral communication test. To investigate raters' behavior in the adaptive testing context, the adaptiveness of the test was first examined by using multilevel ordinal regression analysis. As the between-interviewer variance was more than 10% of the total

score variance based on the intraclass correlation coefficient and the data had a multilevel structure (Nezlek, 2008), the multilevel analysis was first considered to check for any individual interviewers' prompt selection patterns. A random slope cumulative logit model was selected as the best model and fitted by allowing the effect of the explanatory variable, which is the score in the preceding task, on the selection of the following task to vary across all interviewers. Even though about 10% percent of between-interviewer variance was found in the data, the slope of the interviewers' task complexity selection was fairly consistent across interviewers. Thus, mean probability of choosing a task complexity level was used to investigate the adaptiveness of the OECT. Based on the results of the multilevel ordinal regression models, the adaptiveness of the OECT was supported.

However, this adaptiveness of the test would also have created rater- or interviewer-related test score biases. As the principle of adaptive testing is to create a testing context in which the difficulties of test items match the ability of the test takers (van der Linden & Glas, 2010), it can be argued that interviewers in the OECT asked questions adaptively depending on the proficiency level of the test takers. This adaptiveness of the OECT suggests, as the test was designed to address, that the test could have elicited enough ratable oral communication samples from the test takers. However, the result of the partial credit model of many-facet Rasch measurement (MFRM) analysis with the operational OECT data showed that the difficulty levels of the prompts with *low, mid,* and *high* complexity levels statistically turned out to be *high, mid,* and *low,* respectively. This mismatch of task difficulty, which is calculated with the MFRM model, with task complexity, which is assigned by the OECT committee, suggests that test scores calculated with the MFRM analysis in an adaptive

testing context may include unexplained score variances and the scores are less likely to project test takers' true ability.

The mismatch of task difficulty with task complexity might be due to the following two hypothetical reasons: (a) test takers' performance difference with different complexity tasks or (b) raters' rating scale adjustment depending on task complexity. First, the task difficulty could have been measured differently if test takers' performance with different complexity tasks was different and raters thus had to assign test scores accordingly. For example, the difficulty of the *high* complexity could be measured as *low* if the performance of test takers with *high* complexity prompts was worse than the performance level expected by raters. This hypothesis that test takers' performance could have affected the task difficulty, however, was not supported by the linguistic output analysis. The results of statistical analyses that examined the effect of task complexity on test takers' linguistic outputs and proficiency scores showed that only a few fluency measures, such as *Phonation-Time Ratio, Repairs per AS-Unit,* and *Preparation Time,* were statistically different across test takers' performance in response to different complexity prompts. Moreover, these fluency measures were not even translated into the fluency scores measured by human raters, as was found in Leaper and Riazi (2014). If the performance difference was adequately translated into proficiency scores, the task difficulty of *high* complexity prompts should have been calculated as *high* with the MFRM analysis. Thus, the first hypothesis that test takers' performance might have affected the mismatch between *task difficulty* and *task complexity* is not statistically supported.

Second, raters might have given higher scores than what test takers "deserved" when test takers were given more complex prompts. As the prompts with *high* task complexity

were calculated as *low* difficulty prompts when analyzed with an MFRM model, it is implied that raters adjusted their score assignment. If raters assigned higher scores to test takers with *high* complexity prompts, test takers' weak performance with *high* complexity prompts would not be penalized, thus producing higher scores than their performance deserved when task complexity was not considered. The scores that benefited from raters' rating scale adjustment could have changed the task difficulty in the MFRM analysis. If raters gave more benefits than what test takers were penalized for in their linguistic performance because of the tasks' *high* complexity, the difficulty level of *high* complexity prompts could have been lower than that of *low* complexity ones, which is found in the results of the first sub-study.

Even with the linguistic output analysis and task difficulty analysis using the MFRM model, these findings do not lend full support to the initial hypothesis that raters adjust their score assignment depending on the complexity level of the prompts for the following reasons. First, linguistic output analysis cannot fully explain test takers' performance. The construct of the OECT test includes more than simple linguistic outputs. In addition, the effect of task complexity on test takers' performance is quite complex to understand. For example, more complex tasks are generally known to promote more accurate and complex, but less fluent, linguistic outputs than tasks of lesser complexity (Robinson, 2011a). Thus, a simple comparison of linguistic outputs with different task complexity does not necessarily imply that test takers' performance was the same or different with prompts of different task complexity. Second, as the levels of *task complexity* measured by the OECT committee and *task difficulty* calculated with MFRM were separately measured in the operational data analysis, the comparison of *task difficulty* with *task complexity* would not provide evidence that raters adjusted score assignment more than the test takers deserved. If the score

adjustment by raters with different task complexities is larger than the gap between test takers' performance difference and their ability, test takers are likely to have been assigned higher or lower scores than their true ability. This raters' over-adjustment of scores would have created unexpected task difficulty, which is that *high* complexity prompts become *easy* tasks while *low* complexity ones become *difficult* tasks. The comparison of task complexity with task difficulty in the operational data, however, is in a hypothetical stage, because the initial task complexity was not measured by raters and each rater may have accepted task complexity differently. To address the limitations of the study conducted with the operational data, a study with experimental data was conducted. A study with experimental data may not have generalizable results as one with operational data, but it makes it possible to control specific variables.

Rating scale use with the experimental data (RQ 2)

In order to provide further evidence to the findings of the study which included the operational test data (the first research question), the second research question examined how newly trained raters adjusted their score assignment depending on their understanding of prompt complexity by controlling the variability of task complexity in an adaptive testing context. The main difference between the study with operational data and that with experimental data was that the raters themselves categorized tasks into *high*, *mid*, and *low* complexity in the experimental study while the OECT committee did so in the study with the operational data. This direct categorization of the task complexity by the raters was designed to reduce any biases related to the effect of task complexity on the raters' behavior.

The study with experimental data was analyzed with two partial credit models of MFRM, one with the task complexity and performance scores assigned by the raters in the SA group and the other by the raters in the SNA group. The results of the MFRM analysis with experimental data scored by the raters in the SA group showed that the difficulty level of the *high* complexity prompts declined from 0.20 when raters did not know the prompt (the *Answer Only* rating context) to -0.11 in logits when they got to know the prompt (the *Question and Answer* rating context) ($t[365] = 3.70$, $p = .00$). Furthermore, the result of the MFRM analysis based on the scores assigned by the raters in the SNA group also showed that the difficulty level of the *high* complexity prompt declined from 0.38 in the *Answer Only* context to -0.08 in logits in the *Question and Answer* context ($t[189] = 3.79$, $p = .00$).

A possible factor that could impact the findings is that of the time interval between the *Answer Only* and *Question and Answer* rating contexts. To address this issue, the variation of the task difficulty level of *mid* complexity prompts was used as a reference group to examine the task difficulty variation of tasks with *high* or *low* complexity. As the task difficulty of *mid* complexity prompts did not change from *Answer Only* to *Question and Answer* contexts in both SA and SNA group analyses, it can be argued that the variation of task difficulty was not directly attributed to the time intervals between the *Answer Only* and *Question and Answer* rating contexts. Thus, the results of quantitative analysis lend support to the assumption that raters in both SA and SNA changed their rating severity during their rating of performance-based oral communication audio clips by assigning higher scores than expected to test takers with *high* complexity prompts. If raters did not assign higher than expected scores to test takers who received *high* complexity prompts, the task difficulty of

high complexity prompts probably would not change from the *Answer Only* to *Question and Answer* rating contexts, as was the case for test takers assigned to *mid* complexity prompts.

In an effort to provide further evidence of the degree to which raters adjusted their scores based on prompt difficulty on the OECT, qualitative analyses of raters' thought processes while rating was undertaken. Thus, the mixed methods approach, which included both quantitative and qualitative approaches in a single study (Tashakkori & Creswell, 2007), was adopted in the current study to investigate whether or how raters' understanding of the task complexity changed raters' severity depending on the task complexity. The results of verbal reports and interview analyses verified the inference, which was made with the results of the quantitative analyses, that raters assigned higher scores than expected to test takers with *high* complexity prompts. Even though raters in the SNA group expressed a slightly different opinion about the effect of task complexity on rater severity than those in the SA group, they also commented in the verbal reports about task complexity along with other evaluation categories.

In short, the raters in this adaptive performance-based second language oral communication test were found to have considered task complexity when scoring test takers' performance. This finding was supported by both the quantitative findings of operational and experimental data as well as findings of qualitative rater verbal protocols and interviews.

5.1.2. Discussion of the Main Findings

Task complexity and construct-irrelevant variance

The findings of the current study identified the potential effect of task complexity on rater severity in an adaptive performance-based oral communication test. These findings are

in line with those in the previous studies regarding the effect of interviewer characteristics or behaviors on test scores (O'Loughlin, 2002, 2007; O'Sullivan, 2000; Davis, 2009; Hou, 2006; Morton, Wigglesworth, & Wiliams, 1997; Chartrand & Bargh, 1999; Clark, 2002; Wilson & Wilson, 2005). If task complexity in the current study was replaced with the interviewer variable in previous studies, then the findings of the current study, showing that raters adjusted their rating scale depending on task complexity, would be comparable with those in the past studies. For example, Lazaraton (1996) reported that test takers with less supportive interviewers received better scores, because raters compensated for the lack of support by the interviewers. In addition, McNamara and Lumley (1997) and Morton et al. (1997) found that interviewers' supportivess helps test takers earn higher scores, because it assists in eliciting test takers' true ability. This differential effect of interviewer supportiveness on test scores suggests that raters might have their own rationales for adjusting their rating severity depending on the interviewers' behaviors.

Score variability associated with raters is generally accepted as construct-irrelevant variance, or error, in test scores, which can potentially threaten the fairness of the test (A. Brown, 2005). Thus, many performance-based speaking tests have been designed to minimize the score variability associated with both raters or interviewers (A. Brown, 2005). The minimization of the score variability can also be achieved through statistical methods. Score variability associated with raters is not always problematic as long as it is consistent or systematic within each rater, because the systematic severity difference among raters can be statistically corrected using statistical methods, such as MFRM analysis (Linacre, 2014; McNamara, 1996). However, the statistical correction of the systematic severity difference among raters does not necessarily eliminate all potential threats to the fairness of the test.

Even though the unwanted measurement variability can be adjusted with the statistical model, raters can still be internally consistent, but in idiosyncratic ways, resulting in the construct of the test being measured differently across raters (Weigle, 1998). Thus, the idiosyncratic adjustment of rating severity can unfairly influence test scores when the effect of idiosyncracy on test scores, or random errors, becomes stronger than the consistent score variabilities that can be adjusted by the statistical methods.

Adaptiveness and fairness of the test

The findings of this study from analyses of the experimental data showed that raters in both the SA and SNA group considered task complexity, and both groups of raters systematically adjusted their rating severity when they scored test takers' peformance with *high* complexity prompts. The consistent patterns of rater severity adjustment in the current study are limited to the experimental study context where the three complexity levels (*high, mid,* and *low*) were dissected from one another. As will be discussed in the Methodological Implications section (see Section 5.2.2) of this study, rater severity cannot be successfully adjusted by the MFRM analysis in the adaptive testing context. This statistical limitation on analyzing the potential effect of interviewers' adaptive selection of task complexity on test scores may raise the question of whether the adaptiveness of the test should be maintained, or a more restrictive testing context without interviewers' accommodation to test takers' proficiency should be employed to increase the fairness of the test.

When it comes to the oral communication test, it appears that the adaptiveness of task selection by interviewers differentiates the score variability associated with raters from that of a more static testing context where interviewers do not have much freedom to

accommodate to test takers' proficiency level. Previous studies that examined the score variability associated with interviewers mostly focused on the direct effect of the group variable, which is interviewers' characteristics or behaviors, such as the interviewers' gender (O'Loughlin, 2002, 2007; O'Sullivan, 2000) or helpfulness (Chartrand & Bargh, 1999; Clark, 2002; Wilson & Wilson, 2005), on the test scores. Thus, these studies could find the differential effect of an interviewer variable on test scores. In contrast, the current experimental study investigated how raters adjusted their score assignments when they became aware that there was an interviewer variable (or task complexity). This study was designed not to simply measure the test takers' score difference depending on the task complexity levels, but to investigate whether raters would adjust their rating scale to compensate for the potential effect of task complexity on test takers' performance. In the current study design, if raters noticed any unexpected or unfair effect of task types on test takers' performance, they could address the unexpected effect by adding an additional adjustment to their score assignment.

As many factors, such as raters, interviewers, technology, etc., are involved in oral communication assessment (Ockey & Li, 2015), it is likely to have multiple sources of construct-irrelevant variance in the test scores. Based on the findings of this study, it can be inferred that the total construct-irrelevant variance in the oral communication test scores, which were aggregated from multiple sources, does not necessarily pose threats to the fairness of the test. If raters' rating scale adjustments were considerred separately from other factors in the oral communication assessment model, rater behavior would be considered one of the sources of construct-irrelevant variance in the test scores. However, if raters adjusted their score assignment to address any construct-irrelevant variance caused by the

interviewers' adpative selection of task complexity, then the raters' behaviors could be considered as the *moderator variable* (Hayes, 2018) that adjusted any unfair effect of test adaptivenss on the test scores. Based on the qualitative data analysis of the current study, which is that both raters in the SA and SNA groups considered task complexity during their scoring, it appears that raters adjusted their score assignment not to give an unfair advantage or disadvantage to test takers performing a task at different task complexities.

In short, the adaptiveness of the performance-based oral communication test appears not only to provide a better opportunity for interviewers to elicit more ratable speech samples from test takers, but also to put raters in an environment where they adjust their score assignments based on the task complexity, which is likely to help generate more generalizable and fairer test scores. However, the results of the adaptive performance-based test should be interpreted with caution as long as each rater's idiosyncratic adjustment of rating severity are larger than the consistent score variabilities that can be adjusted by the statistical methods.

5.2. Implications of the Study

The findings of the present study have theoretical, methodological, and practical implications. Theoretical implications concern the interaction between raters and task types in the oral communication assessment model (Ockey & Li, 2015). The methodological implications enhance the understanding of task difficulty in the MFRM model of the adaptive performance-based oral communication test. Practical implications provide suggestions about how the adaptive nature of the OECT should be reconsidered.

5.2.1. Theoretical Implications

In their oral communication assessment model, Ockey and Li (2015) introduced several factors, including raters, rating scales, interlocutors' personal characteristics, speech samples, technology, speaking performance, and test takers' oral communication ability, that have an impact on test scores during the administration of oral communication tests. Among many factors that affect test scores in oral communication assessment, raters have been considered as one of the biggest sources of the construct-irrelevant variance in this form of performance-based assessment (McNamara, 1996). A number of past studies on rater effects mostly focused on raters' perception of interviewer variables, such as interviewers' gender (O'Loughlin, 2002, 2007; O'Sullivan, 2000), language proficiency (Davis, 2009), ethnicity (Hou, 2006), competency (Morton et al., 1997), areas of concern (May, 2011), and helpfulness (Chartrand & Bargh, 1999; Clark, 2002; McNamara & Lumley, 1997; Morton et al., 1997; Wilson & Wilson, 2005), but little attention has been paid to the way raters perceive task type during the administration of the test.

In the original model of assessment of oral communication by Ockey and Li (2015), the effect of task type on raters' interpretation of rating scales is only measured through the indirect effect of prompt difficulty, an aspect of task type on test takers' performance. The findings of the current study provided evidence that prompt difficulty might have an interaction with raters, specifically with how severe they rate a speaking performance. This interaction means that individual raters might differently apply rating scales depending on task complexity. The red arrow from *Task Type* to *Rater* in Figure 5.1, which is drawn based on the findings of the current study, describes how raters adjusted their rating severity based on task complexity. Ockey and Li's (2015) original oral communication model assumes that

the task type only influences test takers' speaking performance, which accordingly affects test takers' scores. On the other hand, the modified model in Figure 5.1 adds an additional variable that indicates the interaction of raters' perception of task type with raters' interpretation of rating scales to assign test scores, thus affecting the difficulty level of the prompts and rater severity. The results of this study provided evidence that there could be an additional relationship between task type and raters in Ockey and Li's (2015) oral communication assessment model.

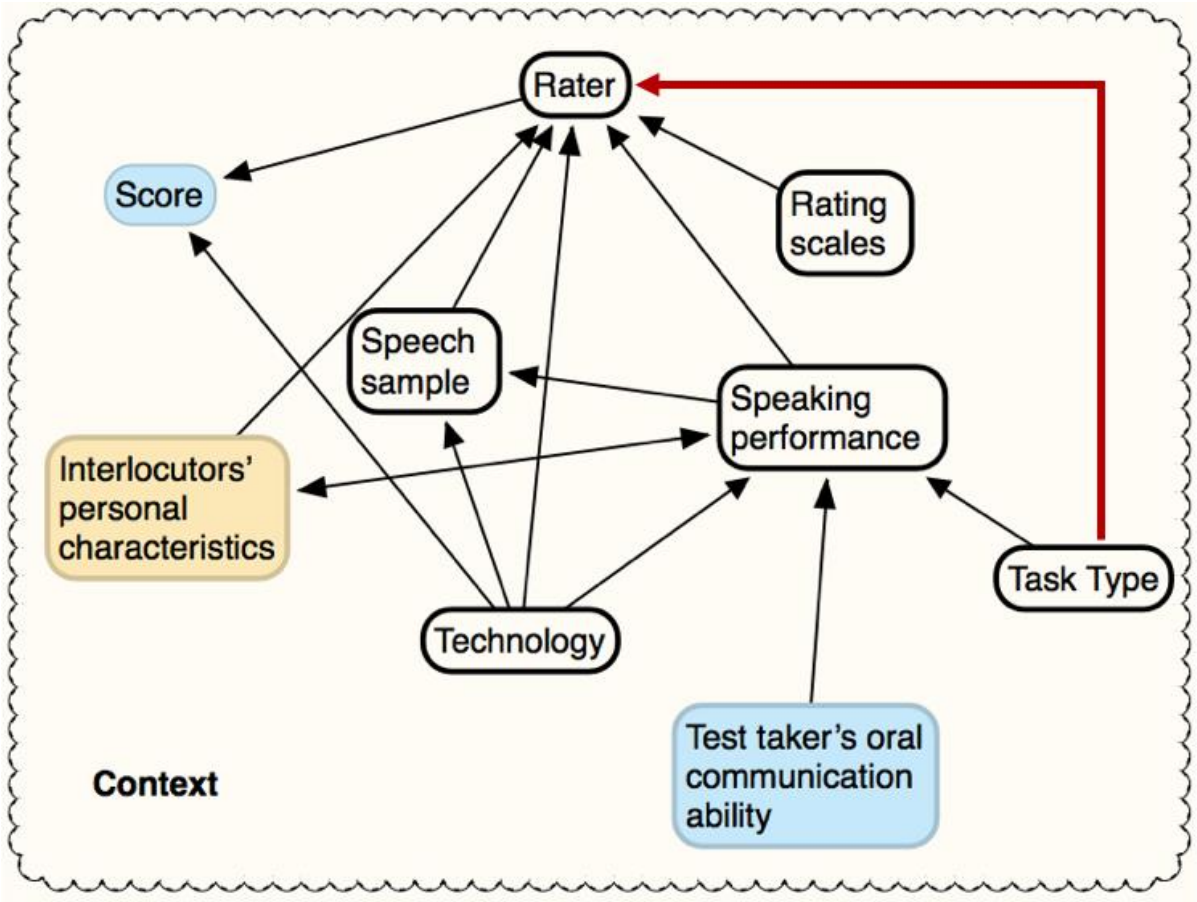


Figure 5.1. The modified model of assessment of oral communication (adapted from Ockey & Li, 2015)

5.2.2. Methodological Implications

MFRM, which was employed for the current study, is a probabilistic measurement model that can calibrate parameters of the model independently of each other (Bond & Fox, 2015; Eckes, 2015; McNamara, 1996). Unlike classical test theory, the MFRM model has the advantage of sample-independent parameter estimation, and item difficulty in the MFRM thus can be independently measured. As was found in the results of the MFRM analysis in the experimental study (Sub-Study 2), however, item difficulty, or task difficulty, was not consistent from the *Answer Only* to *Question and Answer* rating contexts. In addition, the task difficulty, measured with MFRM, was different from task complexity, which was categorized by the OECT committee, when the prompts were administered adaptively to match test takers' ability. This adaptive prompt administration and variability of raters' interpretation of task complexity might have distorted the calculation of item difficulty in the MFRM model.

Even though unexpected item difficulty was found in the current study, the infit and outfit mean-squares were all within the acceptable range (0.5-1.5) (Linacre, 2014), indicating the estimation with the proposed model was successful. The infit or outfit statistics are sensitive to the inlying or outlying observations, respectively (Linacre, 2014), which means the infit or outfit statistics of an item or a rater are calculated in relation to other items' scores or other raters' ratings. If a majority of raters in the performance-based test are systematically lenient towards the test takers' performance with a certain complexity prompt, the infit and outfit statistics can still be within the acceptable range while the model produces biased test scores with different complexity prompts.

The current study raised the question that different task complexity affects not only the test takers' performance, but also raters' severity. The fit indices in the MFRM model should be interpreted with caution.

5.2.3. Practical Implications

The current study provided empirical evidence that the adaptive nature of the OECT should be reconsidered for the following reason. The task complexity in the OECT test influenced not only test takers' performance, as was reported in other studies (Ellis, 2009; Jackson & Suethanapornkul, 2013; Robinson, 2011a; Skehan, 2009), but also raters' rating scale use. The effect of task complexity on raters' rating scale use presumably distorted the task difficulty in the operational OECT data analysis. In typical adaptive testing, item difficulty is calculated before the administration of the test and a test taker is given items until the difficulties of the test items match the ability of the test taker (van der Linden & Glas, 2010). If the OECT follows item selection and ability estimation in typical adaptive testing, raters are required to assign scores solely based on test takers' performance to calculate the "pure" item difficulty.

In the speaking test, however, task difficulty cannot be soly defined by test takers' performance on the task (Fulcher & Reiter, 2003). As was previously discussed, raters in the OECT adjusted their severity during the administration of the test by interacting with other factors in the speeking test. When a test taker was given tasks with task difficulties that were higher or lower than the ability of the test taker, mostly in the first or second tasks of the OECT, the gap between the task difficulty and the test taker ability was reduced not only with the selection of the following task, which is the way for which adaptive testing is

typically designed, but also with raters' severity adjustment during the administration of the task. This raters' severity adjustment made the adaptive nature of the OECT different than that of the typical adaptive testing, thus making the calculation of the parameters of the OECT estimated by the MFRM model not solid.

The raters' adjustment of their rating severity, however, can also be interpreted as an essential element in an adaptive performance-based oral communication test. As was discussed with Ockey and Li's (2015) oral communication assessment model, a plethora of factors are involved in the speaking assessments, and their interaction with one another can create multiple sources of construct-irrelevant variance in test scores. If raters' rating scale adjustment can mitigate any of these unwanted score variabilities, then the findings of this study can be included in the rater training, which makes raters aware of what they should do when they encounter with any potential sources of biases during the administration of the speaking test.

On the other hand, when rater training fails to train raters to ideally adjust their rating severity to mitigate the undesired construct-irrelevant variance in test scores, the adaptiveness of the speaking test should be reconsidered. In this case, a possible way to minimize the task complexity-related, construct-irrelevant variance of the test score would be to remove the adaptiveness of the item selection from the OECT by giving three questions of three different complexity to every test taker (if there are only three levels of task complexity) and requiring raters to focus only on test takers' linguistic performance. For example, both *high* and *low* proficiency test takers would be equally given tasks of *low, mid,* and *high* complexity, which makes at least one task matches the ability of the test taker. As raters are asked to focus only on test takers' linguistic outputs, their rating severity will be

less likely to be affected by the task complexity. The final score of a test taker can be assigned following the test taker's score of the task with which he/she performed best. Assuming that a test taker can perform best and have a "fair" score with the task whose task difficulty matches the ability of the test taker, the fair score may best show the ability of the test taker. The average score of three tasks can also be used as the final score. The average score, however, is less stable than the fair score in that the task whose task difficulty does not match the ability of the test taker would make the test taker produce an inadequate amount of ratable speech (when the task is too difficult) or less complex utterances (when the task is too easy).

5.3. Suggestions for Future Research

A number of suggestions for future research have been made. First, the findings in the study with operational data were based on either the first-round data or the aggregated data from all three rounds. The MFRM analysis with the operational data was based on the assumption that the effect of task complexity on rater severity does not change from the first round to the last round of rating. For example, if raters considered task complexity more or less important in the first round than in the following rounds, the estimation of task difficulty in the MFRM analyses would have been different. Thus, it would be meaningful to investigate whether the round of rating affects raters' perception of task complexity, thus resulting in a different estimation of task complexity.

Second, task complexity was categorized into three levels, *high, mid,* and *low,* but more than 50 prompts were used in the current study. Even though *low* complexity prompts (e.g., *How did you rent your apartment here in Ames?*) appeared easier than *high* complexity

prompts (e.g., *Some people believe that dreams have meaning. Why do people try to interpret dreams?*), there could be some *low* complexity prompts that can be perceived more difficult by test takers. For example, if test takers were new to the U.S., then the questions that require test takers to think about their lives in the U.S. would place them in a hypothetical context, thus making test takers perceive *low* complexity prompts as more difficult. Therefore, it would be meaningful to conduct research with a smaller number of prompts of which the task complexity was screened by coders who are from the same background (e.g., English fluency, cultural background) as the test takers.

The last suggestion to be offered is to study how to decide the final test scores in an adaptive performance-based oral communication test. As was discussed in the Results chapter, test scores calculated with the MFRM analysis could have included task complexity-related, construct-irrelevant variance. It is generally possible to control construct-irrelevant variances in an experimental study condition, but it is almost impossible to secure such conditions in the real testing context. Thus, further studies examining the effect of task complexity not only on rater severity, but also on the final test scores will expand understanding of how factors in the oral communication assessment model (Ockey & Li, 2015) interact with one another and will enable enhanced interpretation of test scores in an adaptive performance-based test.

5.4. Concluding Remarks

Among various factors affecting test scores in an oral communication assessment (Ockey & Li, 2015), human participants in oral communication assessments are one of the major causes of construct-irrelevant variance (McNamara, 1996). Many efforts have been

extended to minimize the score variability associated with raters and interviewers (A. Brown, 1995), but these efforts have made performance-based oral communication assessment reflect less of the target language use domain by controlling interviewers' behaviors.

This study investigated the effect of task complexity on rater severity in an adaptive performance-based oral communication test. The findings of this study suggest that raters may adjust their rating severity depending on how they perceive task complexity. The adjustment of raters' rating scale use may offset the quality of typical adaptive testing, matching the item difficulty with the ability of the test taker (van der Linden & Glas, 2010), in the adaptive oral communication test. Thus, more foundational research that builds upon the current study and explores the effect of interviewers' adaptive behaviors on test scores is needed to fully understand the variability of test scores and to minimize construct-irrelevant variance in performance-based oral communication assessments.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Arbuckle, J. L. (2013). *IBM SPSS Amos 22 user's guide*. Crawfordville, FL: IBM.

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful langauge tests*. Oxford: Oxford University Press.

Bailey, K. (1983). Foreign Teaching Assistants at U.S. Universities: Problems in Interaction and Communication. *TESOL Quarterly, 17*(2), 308-310.

Barkaoui, K. (2013). Using multilevel modeling in language assessment research: A conceptual introduction. *Language Assessment Quarterly, 10*(3), 241-273.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 45-80). Norwood, NJ: Ablex.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters: Design, innovation, and discovery* (Vol. 2). New York: Wiley-Interscience.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Peter Lang.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413-425). New York: Routledge.

Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 37-62). Cambridge: Cambridge University Press.

Brown, A., & McNamara, T. F. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly, 38*(3), 524-538.

Brown, G., Anderson, A., Shillcock, R., & Yule, G. (1984). *Teaching talk: Strategies for production and assessment*. Cambridge: Cambridge University Press.

Browne, W. J. (2015). *MCMC estimation in MLwiN: Version 2.32* Retrieved from http://www.bris.ac.uk/cmm/media/software/mlwin/downloads/manuals/2-32/mcmc-web.pdf

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

Chalhoub Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second langauge acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing anguage through computer technology*. Cambridge: Cambridge University Press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893-910.

Clark, H. H. (2002). Speaking in time. *Speech Communication, 36*, 5-13.

Cohen, A. (1998). *Strategies in learning and using a second language*. London: Longman.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155-159.

Cotos, E. (2014). *Oral English Certification Test (OECT): Rater manual*. Ames, IA: Iowa State University.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). CA: Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281-302.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197-222.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*(4), 309-316.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guildford.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385-390.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170–177.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main: Peter Lang.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*(4), 474-509.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: The MIT Press.

Farnsworth, T. (2013). Assessing the oral English abilities of international teaching assistants in the USA. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 471-483). Hoboken, NJ: John Wiley & Sons, Inc.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage publications.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Pearson.

Fulcher, G., & Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344.

Ginther, A. (2003). International teaching assistant testing: Policies and methods. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (pp. 57-84). Washington: NAFSA.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York: Guilford Publications.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communicatin Methods and Measures, 1*(1), 77-89.

Hou, Y.-c. (2006). *A cross-cultural study of the perception of apology—Effect of contextual factors, exposure to the target language, interlocutor ethnicity and task language.* (Unpublished master's thesis), National Sun Yat-sen University, Taiwan.

Hout, R. v., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). Cambridge: Cambridge University Press.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological bulletin, 105*(2), 301-307.

Jackson, D. O., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*(2), 330-367.

Karavas, E., & Delieza, X. (2009). On site observation of KPG oral examiners: Implications for oral examiner training and evaluation. *Journal of Applied Language Studies, 3*(1), 51-77.

Katz, A., & Gottlieb, M. (2013). Assessment in the classroom. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-8). Blackwell Publishing.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment.* (Unpublished doctoral dissertation), Columbia University.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239-261.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81-112.

Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 39-66). Philadelphia: John Benjamins.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Education and Psychological Measurement, 30*, 61-70.

Krippendorff, K. (2004a). *Content analysis: An introduction to its methodology* (2nd ed.). London: Sage Publications.

Krippendorff, K. (2004b). Reliability in content analysis. *Human Communication Research, 30*(3), 411-433.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. CA: SAGE Publications

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*(2), 151-172.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing, 31*(2), 177-204.

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *48*(4), 399-418.

Leech, G. N., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. New York: Taylor & Francis.

Lennon, P. (2006). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Linacre, J. M. (2012). Facets tutorial. Retrieved from http://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2014). *A user's guide to FACETS* Retrieved from https://www.winsteps.com/a/Facets-ManualPDF.zip

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*(4), 483-490.

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design*. New York: Routledge.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). *The Stanford CoreNLP natural language processing toolkit*. Paper presented at the Association for Computational Linguistics (ACL) System Demonstrations.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140-156.

Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 141-173). Amsterdam: John Benjamins.

Morton, J., Wigglesworth, G., & Wiliams, D. (1997). Approaches to the evaluation of the interviewer performance in oral interaction tests. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in English language test design and delivery.* (pp. 175-196). Sidney: National Centre for English Language Teaching and Research.

Myford, C., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement, 5*(2), 189-227.

Nakatsuhara, F. (2011). The relationship between test-takers' listening proficiency and their performance on the IELTS speaking test. *IELTS Research Reports Volume 12, 2011*, 1-50.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*(2), 842-860.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. London: SAGE.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

O'Loughlin, K. (2007). An investigation into the role of gender in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers Research in speaking and writing assessment* (pp. 63-97). Cambridge: Cambridge University Press.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373-386.

Ockey, G. J., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value, 7*(1), 1-21.

Piston Software. (2017). *Direct WAV MP3 Splitter*. Retrieved from http://www.pistonsoft.com/mp3-splitter.html

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford: Oxford University Press.

Qualtrics. (2018). Qualtrics (Version November 2018). Provo, UT, USA: Qualtrics. Retrieved from https://www.qualtrics.com/

Rasbash, J., Browne, W., Healy, M., & Cameron, B. (2015). *MLwiN Version 2.36*. Centre for Multilevel Modelling, University of Bristol.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

Révész, A., & Gurzynski-Weiss, L. (2016). Teachers' perspectives on second language task difficulty: Insights from think-alouds and eye-tracking. *Annual Review of Applied Linguistics, 36*, 182-204.

Rey, D., & Neuhäuser, M. (2011). Wilcoxon-Signed-Rank test. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1658-1659). Berlin, Heidelberg: Springer Berlin Heidelberg.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14*(4), 423-441.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). New York: Cambridge University Press.

Robinson, P. (2011a). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of langauge learning and performance* (pp. 3-37). Amsterdam: John Benjamins.

Robinson, P. (2011b). Task-based language learning: A review of issues. *Language Learning, 61*(Suppl. 1), 1-36.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing, 9*(2), 173-185.

Ross, S. (2012). Claims, evidence, and inference in performance assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 223-233). New York: Routledge.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin, 88*(2), 413-428.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. London: Sage.

Skehan, P. (1998). *A cognitive approach to language learning*. Boston: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning and testing* (pp. 167-185). London: Longman.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-211.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage Publications.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education Limited.

Tarone, E. (1998). Research on interlanguage variation: Implications for language testing. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 71-89). Cambridge: Cambridge University Press.

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research, 1*(1), 3-7.

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *IRAL, 40*(2), 117-150.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111.

van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. NY: Springer.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411-440.

Van Moere, A. (2013). Raters and ratings. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1358-1374). Hoboken, NJ: John Wiley & Sons, Inc.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weir, C. J., O'Sullivan, B., & Horal, T. (2006). 5. Exploring difficulty in Speaking tasks. *IELTS Research Report, 6*, 1-42.

Welch, B. L. (1947). The generalization of `Student's' problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35.

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment, 21*(6), 765-774.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review, 12*(6), 957-968.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35-51.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA press.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501-527.

Zhang, Y., & Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.

APPENDIX A. OECT SCORING RUBRIC

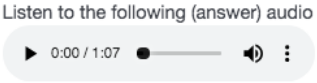
	Band (score) (New & Analytic)	Functional Competency	(Pace &) Fluency	Lexico-grammar	Pronunciation
LEVEL 1	**Excellent (28-30)** (13) Communication is like that of an educated N-American native speaker; always fluent & effective; no effort needed to understand.	Support arguments, hypothesize, discuss in detail; highly competent to convey ideas on familiar & unfamiliar, concrete & abstract topics & to handle complicated communicative tasks in all situations.	Native-like delivery	Native-like with sophisticated, appropriate, & precise vocabulary & grammar.	Native-like pronunciation with only very slight "foreign" accent.
	Very strong (25-27) (11-12) Communication is fairly close to that of an educated N-American native speaker; always fluent & effective; little effort needed to understand.	Support arguments, hypothesize, discuss in detail; sustain very strong but not excellent performance; highly competent to convey ideas on familiar & unfamiliar, concrete & simpler abstract topics & to handle complex communicative tasks in most situations.	Effective pace; smooth delivery & fluency, with good use of English rhythm & focal stress to highlight meaning.	Rich vocabulary & accurate grammar; a few unusual expressions or minor problems possible.	Pronunciation highly intelligible; may have a few consistent or minor problems.
	Strong (23-24) (9-10) Communication is generally effective; fluent most of the time; performance slightly weakens with more complicated topics and tasks; little effort needed to understand.	Somewhat support arguments, hypothesize, discuss in less detail; competent to convey ideas on concrete, familiar topics & to handle communicative tasks in many formal and informal situations; linguistic performance slightly weakens when handling abstract, unfamiliar topics or performing more complicated tasks.	Delivery at a fair pace; fluency generally smooth & with good rhythm.	Adequate, but not stellar vocabulary; occasional or slight problems with expression or lexical & grammatical forms.	Pronunciation generally intelligible; may have minor problems of interference from accent.
LEVEL 2	**Adequate (21-22)** (7-8) Communication is fairly effective; fluent most of the time; cannot sustain performance with more complicated topics and tasks; is able to compensate for the limited aspects of communication; some effort needed to understand.	Explain, narrate, describe, compare; fairly competent to convey ideas on concrete, familiar topics & handle unsophisticated tasks in many formal and informal situations; linguistic performance noticeably weakens when handling abstract, unfamiliar topics or performing more complicated tasks.	Delivery at a fair pace; fluency often smooth & with good rhythm, but occasionally choppy or too even.	Good vocabulary; good but inconsistent use of all time frames; noticeable but not serious problems with expression or lexical & grammatical forms.	Pronunciation generally intelligible; some words may be unintelligible due to pronunciation errors or accent.
LEVEL 3	**Limited (19-20)** (5-6) Communication somewhat effective; speaker expresses ideas freely, but has problems that impede communication; more effort needed to understand.	Explain, narrate, describe, compare in simple ways, maintain conversation; able to convey ideas on basic & concrete topics of personal relevance in informal and few formal situations; can occasionally perform functions of Level 2 but unable to sustain performance.	Delivery may be overly slow or fast; fluency with choppy flow, or very even rhythm, or word by word at times, or inappropriate intonation patterns.	Fair vocabulary; fair use of time frames & expression; sentence structure problems.	A few words or phrases unintelligible due to frequent or consistent pronunciation errors or due to strong accent.
	Very limited (17-18) (3-4) Some communication takes place, but speaker struggles to express ideas and/or has significant communication problems that make it difficult for listener to understand.	Barely create with the language to narrate, describe & ask/answer q-ns; able but struggle to convey ideas on concrete predictable topics in simple situations related to self; able to understand the task but unable to address the requirements of the task.	Pace is often slow because of halting fluency with hesitations and/or repetitions.	Limited vocabulary; lack of grammar control; frequent and/or serious errors.	Pronunciation errors, poor enunciation, or a strong accent may make some ideas unintelligible.
LEVEL 4	**Poor (12-16)** (1-2) Communication generally ineffective; multiple significant problems exhibited; much effort needed to understand.	Little to no functional ability; able to provide basic information & respond to simple questions/requests but use language reactively; often unable to understand the task.	Inappropriate pace may significantly interfere; broken fluency because of pauses, hesitations, & false starts.	Poor to almost lacking control of vocabulary & grammar.	Pronunciation may make many ideas unintelligible.
	Not competent (0-11) (0) Totally ineffective communication; listener can only catch a few words.	No functional ability, able to imitate & re-cycle interlocutor's words; unable to understand the task.	Not fluent; pace of delivery severely interferes.	Lack of vocabulary & grammar.	Pronunciation may make most ideas unintelligible.

APPENDIX B. SCORING PAGE FOR RATERS IN QUALTRICS

Listen to the following (question) audio file, and indicate how difficult this question would be to international graduate students who just arrived in the U.S.

	Extremely easy	Very easy	Somewhat easy	Somewhat difficult	Very difficult	Extremely difficult
Question Difficulty	○	○	○	○	○	○

Listen to the following (answer) audio file and evaluate his/her speaking proficiency based on the given scoring rubric below.



	Not Competent 0	- Poor 1	+ Poor 2	- Very Limited 3	+ Very Limited 4	- Limited 5	+ Limited 6	- Adequate 7	+ Adequate 8	- Strong 9	+ Strong 10	- Very Strong 11	+ Very Strong 12	Excellent 13
Holistic	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Band (score)
Excellent (13) Communication is like that of an educated N-American native speaker; always fluent & effective; no effort needed to understand.
Very strong (11-12) Communication is fairly close to that of an educated N-American native speaker; always fluent & effective; little effort needed to understand.
Strong (9-10) Communication is generally effective; fluent most of the time; performance slightly weakens with more complicated topics and tasks; little effort needed to understand.
Adequate (7-8) Communication is fairly effective; fluent most of the time; cannot sustain performance with more complicated topics and tasks; is able to compensate for the limited aspects of communication; some effort needed to understand.
Limited (5-6) Communication somewhat effective; speaker expresses ideas freely, but has problems that impede communication; more effort needed to understand.
Very limited (3-4) Some communication takes place, but speaker struggles to express ideas and/or has significant communication problems that make it difficult for listener to understand.
Poor (1-2) Communication generally ineffective; multiple significant problems exhibited; much effort needed to understand.
Not competent (0) Totally ineffective communication; listener can only catch a few words.

APPENDIX C. RETROSPECTIVE VERBAL REPORT GUIDELINE

(adapted from Ericsson & Simon, 1993, p. 378 and Green, 1998)

Introduction:

In this task, I am interested in what you think about when you rate the audio clips using the holistic rating scale. In order to do this, I am going to ask you to TALK ALOUD after listening to the audio clip. What I mean by TALK ALOUD is that I want you to tell me everything you are thinking about the audio clips in terms of your grading based on the evaluation rubric (Appendix A). To successfully complete this task, I would like you to talk aloud constantly until you do not have anything to add. I don't want you to plan out what to say or try to explain to me what you are saying. If you are silent for any long period of time, I will ask you to talk.

Warm-up:

Good, now let's begin with a practice problem. I want you to verbalize your thinking process for the following question.

> "A bottle of wine costs $5. The wine costs $4.50 more than the bottle. How much does the bottle cost?"

Now you will hear one minute of an audio clip. Please evaluate its proficiency level.

> *<play the sample audio clip> (60 seconds)*
>
> What is the overall proficiency score of this file?

Now please verbalize your cognitive processing regarding the proficiency level and tasks of the following audio clip based on the rubric. I will play 20 seconds for each clip.

> *<play the first 20 seconds>*
>
> *<wait for the response. If no response, then say>* Any idea about this audio clip?
>
> *<play the second 20 seconds>*
>
> *<wait for the response. If no response, then say>* Any idea about this audio clip?
>
> *<play the third 20 seconds>*
>
> *<wait for the response. If no response, then say>* Any idea about this audio clip?

Main session:

Now let's move on to the main session. First, let's listen to the following audio clips and evaluate their overall proficiency score based on the given scoring rubric.

<Play audio clip NO. 1 (60 seconds)>

What is the overall proficiency score of this file?

<Play audio clip NO. 2 (60 seconds)>

What is the overall proficiency score of this file?

Now please verbalize your cognitive processing regarding the proficiency level and tasks of the following audio clip based on the rubric. I will play 20 seconds for each clip.

<play the first 20 seconds of audio clip No. 1>

<wait for the response. If no response, then say> Any idea about this audio clip?

<play the second 20 seconds of audio clip No. 1>

<wait for the response. If no response, then say> Any idea about this audio clip?

<play the third 20 seconds of audio clip No. 1>

<wait for the response. If no response, then say> Any idea about this audio clip?

<play the first 20 seconds of audio clip No. 2>

<wait for the response. If no response, then say> Any idea about this audio clip?

<play the second 20 seconds of audio clip No. 2>

<wait for the response. If no response, then say> Any idea about this audio clip?

<play the third 20 seconds of audio clip No. 2>

<wait for the response. If no response, then say> Any idea about this audio clip?

Repeat the main session with audio clips No. 3 & 4.

<No. 3 & 4 includes interviewer's prompt; follow same procedure but add following questions to 20 seconds split play; after the first 20 seconds>

any comment on the prompt?

Repeat the main session with audio clips No. 5 & 6.

<same as with No. 1 & 2>

Repeat the main session with audio clips No. 7 & 8.

<No. 7 & 8 includes interviewer's prompt; follow same procedure, but add following questions to 20 second split play; after the first 20 seconds>

Any comment on the prompt?

APPENDIX D. SEMI-STRUCTURED INTERVIEW FOR RATERS

Thank you for participating in the rater's perception study. This questionnaire asks about your personal perceptions about how you behave as a rater in the performance-based oral communication test.

1. **(As a rater)** When you grade a test taker's performance score in each task, which evaluation category in the scoring rubric do you consider most? Please order them in terms of their importance and assign the proportion of their importance.

Category	**Functional competency**	**Fluency**	**Lexico-grammar**	**Pronunciation**
Order (1^{st} – 4^{th})				
Proportion (%)				

(Please explain why: _____)

2. **Task Complexity**
   a. I apply the same evaluation criteria for the holistic score when test takers must deal with topics with different complexity (or difficulty) (e.g., *high* vs. *low* complexity).

Strongly Disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly Agree**
1	2	3	4	5

(Please explain why: _____)

   b. I focus on a test taker's linguistic outputs when I grade his/her performance without considering the task complexity (or difficulty).

Strongly Disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly Agree**
1	2	3	4	5

(Please explain why: _____)

   c. I take the complexity (or difficulty) into consideration when I grade a test taker's performance (e.g. slow speed with difficult items vs. fast speed with easy items).

Strongly Disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly Agree**
1	2	3	4	5

(Please explain why: _____)

3. **(Perception of Interviewer's performance)** Do you take the following interviewers' behaviors into consideration when you evaluate a test taker's performance in a task?

   a. I consider the interviewer's inappropriate topic (or prompt content) choice when I grade a test taker's performance (e.g., topics culturally challenging to understand).

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
1	2	3	4	5

(Please explain why: _____)

   b. I consider the interviewer's unhelpful (or inappropriate) questions when I grade a test taker's performance.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
1	2	3	4	5

(Please explain why: _____)

   c. I consider the interviewer's selection of closed follow-up questions when I grade a test taker's performance (e.g., asking simple yes/no questions).

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
1	2	3	4	5

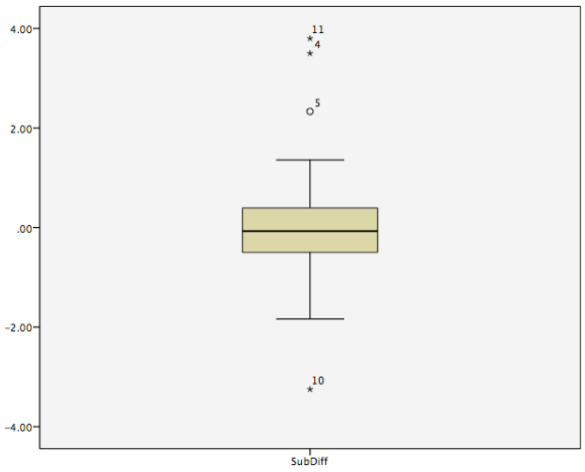
(Please explain why: _____)

   d. I consider the interviewer's over-domination of the conversation when I grade a test taker's performance (e.g., not giving enough time to test takers).

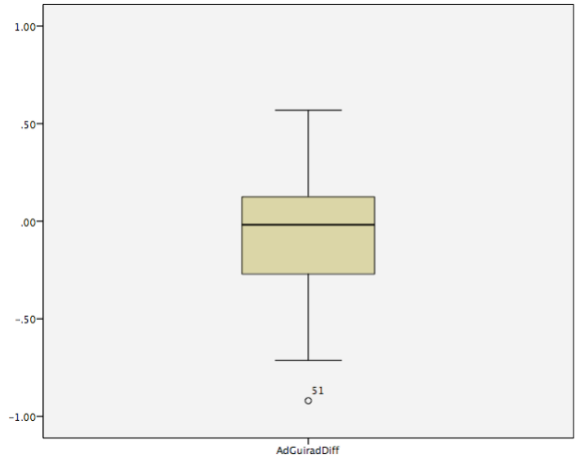
Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
1	2	3	4	5

(Please explain why: _____)

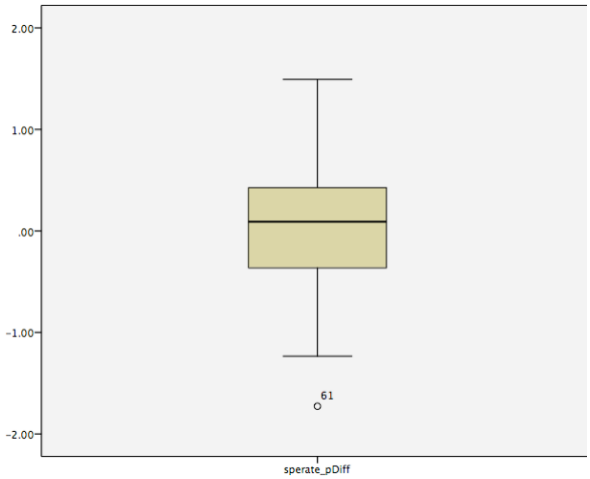
APPENDIX E. SYMMETRICAL DISTRIBUTION OF LINGUISTIC FEATURE DIFFERENCES



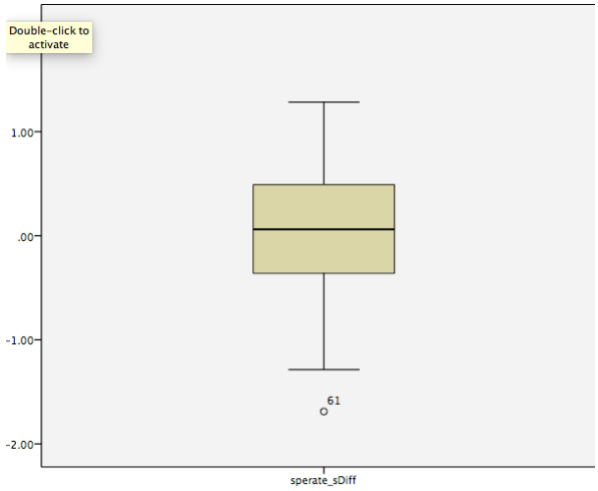
(Subordinate Index Difference)



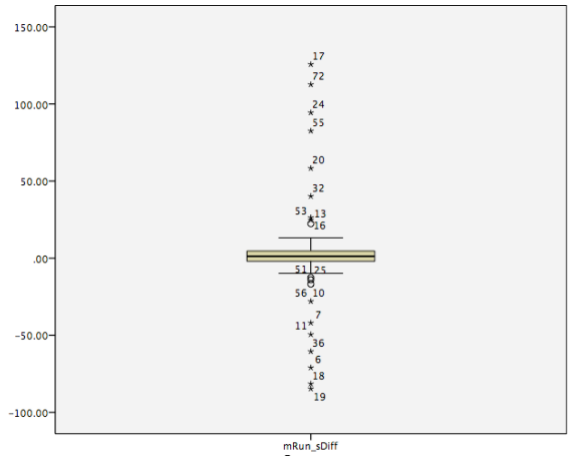
(Guiraud Advanced 1000 Difference)



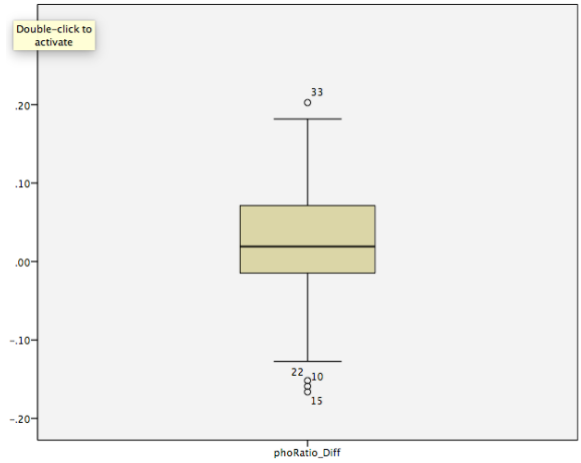
(Pruned Speech Rate Difference)



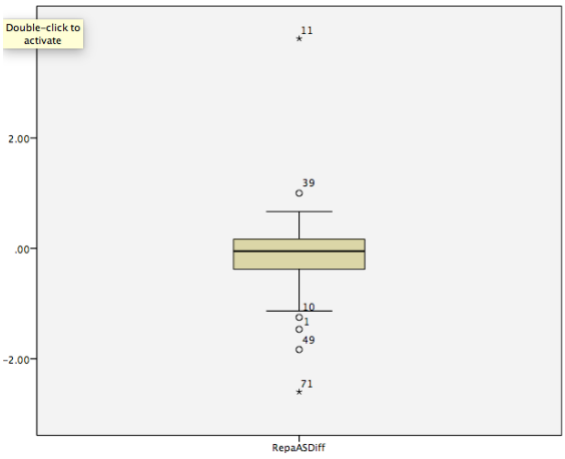
(Unpruned Speech Rate Difference)



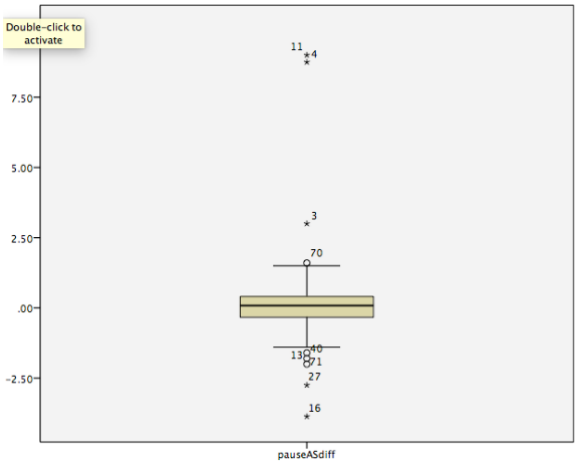
(Mean Length of Runs Difference)



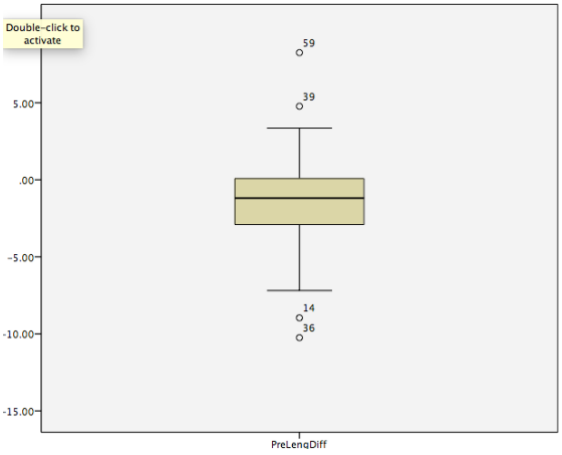
(Phonation Time Ratio Difference)



(Repairs per AS-unit Difference)



(Filled Pauses per AS-unit Difference)



(Preparation Time Difference)

APPENDIX F. INSTITUTIONAL REVIEW BOARD APPROVAL

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 03/22/2018

To: Yongkook Won Gary Ockey

From: Office for Responsible Research

Title: **Interplay between Task Complexity and Interlocuters/Raters in a Semi-Adaptive Performance-based Second Language Oral Communication Test**

IRB ID: **16-050**

Submission Type: Continuing Review & Modification **Review Type:** Expedited

Approval Date: 03/22/2018 **Date for Continuing Review: 03/21/2020**

The project referenced above has received approval from the Institutional Review Board (IRB) at Iowa State University according to the dates shown above. Please refer to the IRB ID number shown above in all correspondence regarding this study.

To ensure compliance with federal regulations (45 CFR 46 & 21 CFR 56), please be sure to:

- **Use only the approved study materials** in your research, including the **recruitment materials and informed consent documents that have the IRB approval stamp.**
- **Retain signed informed consent documents for 3 years after the close of the study,** when documented consent is required.
- **Obtain IRB approval prior to implementing any changes** to the study.
- **Inform the IRB if the Principal Investigator and/or Supervising Investigator end their role or involvement with the project** with sufficient time to allow an alternate PI/Supervising Investigator to assume oversight responsibility. Projects must have an eligible PI to remain open.
- **Immediately inform the IRB of (1) all serious and/or unexpected adverse experiences** involving risks to subjects or others; and (2) **any other unanticipated problems involving risks** to subjects or others.
- **Stop all human subjects research activity if IRB approval lapses,** unless continuation is necessary to prevent harm to research participants. Human subjects research activity can resume once IRB approval is re-established.
- **Submit an application for Continuing Review** at least three to four weeks prior to the **date for continuing review** as noted above to provide sufficient time for the IRB to review and approve continuation of the study. We will send a courtesy reminder as this date approaches.

IRB 03/2018

- Please be aware that IRB approval means that you have met the requirements of federal regulations and ISU policies governing human subjects research. **Approval from other entities may also be needed.** For example, access to data from private records (e.g. student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **IRB approval in no way implies or guarantees that permission from these other entities will be granted.**

- Please be advised that your research study may be subject to **post-approval monitoring by Iowa State University's Office for Responsible Research**. In some cases, it may also be subject to formal audit or inspection by federal agencies and study sponsors.

- Upon completion of the project, transfer of IRB oversight to another IRB, or departure of the PI and/or Supervising Investigator, please initiate a Project Closure to officially close the project. For information on instances when a study may be closed, please refer to the IRB Study Closure Policy.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.